From Department of Medical Biochemistry and Biophysics
Karolinska Institutet, Stockholm, Sweden

APPLICATIONS OF GENOMIC TOOLS TO DECODE GENOME FUNCTION

Jilin Zhang

Karolinska Institutet

Stockholm 2020

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB

ISBN 978-91-7831-791-2

Karolinska Institutet

Institutionen för medicinsk biokemi och biofysik

Applications of genomic tools to decode genome function

AKADEMISK AVHANDLING
som för avläggande av medicine doktorsexamen vid Karolinska Institutet offentligen försvaras i Biomedicum D0320, Solnavägen 9

Måndagen den 18:e maj 2020, kl 14.00

av

Jilin Zhang

Huvudhandledare:
Professor Jussi Taipale
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Functional Genomics

Bihandledare:
Dr. Minna Taipale
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Functional Genomics

Dr. Bernhard Schmierer
Karolinska Institutet
Department of Medical Biochemistry and Biophysics

Fakultetsopponent:
Assistant Professor Claudia Kutter
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Betygsnämnd:
Associate Professor Carsten Daub
Karolinska Institutet
Department of Biosciences and Nutrition

Professor Björn Högberg
Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Biomaterials

Professor Ann-Christine Syvänen
Uppsala Universitet
Department of Medical Sciences

Stockholm 2020

献给我的祖父

To my grandfather

Till min farfar

ABSTRACT

The revolution in sequencing technologies has greatly advanced our understanding of genomes. Many regulatory elements lacking protein-coding ability, such as long non-coding RNAs have been identified and characterized in different biological contexts. However, functional interrogation of these non-coding elements remains to be challenging when it comes to resolving the relationships between genotypes and phenotypes. To elucidate the functional roles of regulatory elements encoded in the genome and further to deconvolute the evolutionary history of chromosomes, I developed new informatics tools/strategies and combined them with existing computational tools to analyse the genomic data.

In study I, a bioinformatics strategy was developed and implemented to identify sex-linked sequences and to recover the genes from a set of recently available avian genomes. The analysis of molecular signatures on sex chromosomes across species has described the unique evolutionary trajectories in avian genomes for the first time.

In study II, a novel *in vitro* assay was applied to determine the binding specificities of human RNA binding proteins. By searching for the potential enrichment of their binding sites in the human genome with a newly implemented tool, the essential roles of RBPs involved in many RNA metabolic procedures have been reinforced.

In study III, the unique molecular identifiers were incorporated into the loss-of-function study in CRISPR/Cas9 based pooled screening. I implemented the analytical tools to interpret the data, which has immensely extended the power of pooled screening by allowing to trace phenotypes of individual cell lineages.

In study IV, sequence conservation information contributed by comparative genomics has been integrated to indicate the functional significance of enhancers upstream of the oncogene *Myc*, which, however, counter-intuitively did not show obvious physiological consequences after knockout.

In summary, four studies were conducted to dissect the functionality of the genome. Through integrating knowledge from distinct dimensions, we can eventually attempt to unveil the principles that dictate the relationships between genotypes and phenotypes.

LIST OF SCIENTIFIC PAPERS

I. Zhou, Q., **Zhang, J.**, Bachtrog, D., An, N., Huang, Q., Jarvis, E. D., Gilbert, M. T. P., & Zhang, G. (2014). Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science*, *346*(6215), 1246338.

II. Jolma, A., **Zhang, J.**, Mondragón, E., Kivioja, T., Yin, Y., Zhu, F., Morris, Q., Hughes, T. R., Maher, L. J., & Taipale, J. (2019). Binding specificities of human RNA binding proteins towards structured and linear RNA sequences. Manuscript.

III. Schmierer, B., Botla, S. K., **Zhang, J.**, Turunen, M., Kivioja, T., & Taipale, J. (2017). CRISPR/Cas9 screening using unique molecular identifiers. *Molecular Systems Biology*, *13*(10), 945.

IV. Dave, K., Sur, I., Yan, J., **Zhang, J.**, Kaasinen, E., Zhong, F., Blaas, L., Li, X., Kharazi, S., Gustafsson, C., De Paepe, A., Månsson, R., & Taipale, J. (2017). Mice deficient of Myc super-enhancer region reveal differential control mechanism between normal and pathological growth. *eLife*, *6*.

Publications not included in the thesis

Wang, Z., **Zhang, J.**, Xu, X., Witt, C., Deng, Y., & Chen, G. (2019). Phylogeny, transposable element and sex chromosome evolution of the basal lineage of birds. Manuscript.

Zhang, J., Li, C., Zhou, Q., & Zhang, G. (2015). Improving the ostrich genome assembly using optical mapping data. GigaScience, 4, 24.

Zhang, J., Li, J., & Zhou, Q. (2017). Genomic and Transcriptomic Analyses of Avian Sex Chromosomes and Sex-Linked Genes. Methods in Molecular Biology, 1650, 69–85.

CONTENTS

LIST OF ABBREVIATIONS

ALS	Amyotrophic lateral sclerosis
CRISPR/Cas9	Clustered regularly interspaced palindromic repeats/Cas9
C2H2	Cys-Cys-His-His domain
CLIP	Cross-linking immunoprecipitation
dsRNA	double stranded RNA
DNA	Deoxyribonucleic acid
eCLIP	enhanced crosslinking and immunoprecipitation
GO	Gene Ontology
GWAS	Genome-wide association study
hnRNP	Heterogeneous nuclear ribonucleoprotein
iCLIP	individual-nucleotide resolution UV crosslinking and immunoprecipitation
KH	K homology domain
Ma	Million years
MAD	Median of absolute deviation
NGS	Next-generation sequencing
OM	Optical mapping
PCR	Polymerase chain reaction
PAR-CLIP	Photoactivatable ribonucleotide-enhanced crosslinking and immunoprecipitation
PAR	Pseudoautosomal regions
PWM	Position weight matrix
RBFOX1	RNA binding protein Fox-1 homolog
RNAi	RNA interference
RBPs	RNA binding proteins
RBDs	RNA binding domains
RRM	RNA recognition motif
RSL	Random sequence label
RNA	ribonucleotide acid
RISC	RNA induced silencing complex

RIP	Ribonucleoprotein immunoprecipitation
SD	Sex determining
SELEX	Systematic evolution of ligands by exponential enrichment
SA	Sex antagonism
SLM	Stemloop model
SSMD	Strictly standardized mean difference
TF	Transcription factor
TALEN	Transcription activator like effector nuclease
Znf	Zinc finger

1 INTRODUCTION

All the work presented in this thesis resulted from collaborations with other researchers. Experiments and primary data were generated and collected by collaborators, whereas I contributed to the development of computational tools and analytical frameworks for the projects to address specific questions. As not all the included projects have obvious connections to each other, I first briefly introduced the biological background of each study and scientific questions to be addressed. Next, I summarized my major contributions and acknowledged the contribution of others. I ended the thesis with a discussion of the challenges in decoding the genome. Although the four studies addressed different questions in different fields and present non-related topics, the common goal of the work is to disclose the functional roles of the genomic sequences.

Study I is one of my long-term research interests, the application of comparative genomics to investigate the sex chromosome evolution in avian genomes, which continued through my PhD studies encouraged by my supervisor, Jussi Taipale. With recently sequenced genomes, the evolution of molecular signatures on sex chromosomes was systematically assessed. I developed analytical approaches to identify sex-chromosome linked sequences, characterizing the evolutionary trajectories of sex chromosomes, as well as deducing the dynamics of sex-linked genes. Qi Zhou and Guojie Zhang provided the initial data and ideas. Jussi Taipale granted me the freedom to continue this project and provided access to resources.

Study II presents work to analyse the binding specificities of RNA binding proteins (RBPs) using data generated by high-throughput RNA-SELEX. The experiments were conducted by Arttu Jolma based on the experimental procedures set up by Estefania Mondragón. My contribution was to characterize binding profiles, to analyse the properties of identified motifs, and to interpret potential functional roles using bioinformatic strategies.

Study III was published in Molecular Systems Biology. Optimization of CRISPR screens by inclusion of unique molecular identifiers (UMI) was proposed by Jussi Taipale and tested by Bernhard Schmierer and Sandeep Botla. I implemented a pipeline to analyse the complex data obtained from this new experimental method.

Study IV focused on the functional roles of non-coding regulatory sequences. In this work, I mainly used the conservation data generated by comparative genomics to visualize the evolutionary conservation across species to highlight the functional importance of the studied loci.

The revolution of sequencing technologies has greatly advanced our understanding of genomes. With access to the increasingly large volume of genomic data, many computational tools have been developed, and together with developing laboratory approaches, have been used to answer fundamental questions. For instance, tools in comparative genomics, sequence conservation analysis and ancestral sequence inference have been applied to compare genomes from a set of well-selected species to test hypotheses and examine evolutionary consequences. The analyses of 29 vertebrates revealed many interesting features of conserved non-coding loci that are likely involved in regulatory networks[1]. Moreover, the evolutionary patterns of enhancers in 20 mammalian genomes disclosed the functional importance of both conserved non-coding sequences and species-specific non-coding loci[2]. In addition to the advancement of comparative genomics, knockout experiments in mice have provided insights into the functional roles of certain genomic loci[3,4]. The ease of applying CRISPR/Cas9 technology for genome editing further permits efficient functional interrogation[5–7].

However, challenges to dissect the functional roles of non-coding sequences remain to be overcome: Sequences across species are much more dissimilar compared to protein-coding genes; a comprehensive functional characterization of non-coding sequences and regulatory elements is missing, and relationships between genomic loci and phenotypes are highly complex. Only through decoding the functionality encoded in each genomic locus with appropriate tools, can we lay out a functional map to understand genomes.

1.1 EVOLUTION OF SEX CHROMOSOMES

1.1.1 The origin of sex chromosomes

A wide range of eukaryotic species have evolved sex chromosomes for sexual reproduction[8]. In spite of the diversity within metazoan genomes, species of distant phylogenetic positions employ the XY system, such as silkworm and papaya, indicating that recruiting sex chromosomes is an ancient strategy[9]. Nonetheless, sex chromosomes among different systems usually differ dramatically without direct homology between chromosomes. The XY system (male heterogametic) of mammals and ZW system (female heterogametic) of birds suggest that sex chromosomes in these species evolved from different pairs of ancestral chromosomes[10–12]. In addition to the different origins, variation of the responsible sex

determining (SD) genes adopted by different SD systems within relatively close species indicates a rapid turnover of sex chromosomes[13].

The hypothesis that sex chromosomes evolved from a pair of autosomal-like ancestral chromosomes has been reinforced by generating a massive volume of whole-genome sequences from many distant species and comparing genomes from various phylogenetic clades[14,15] (**Figure 1**).

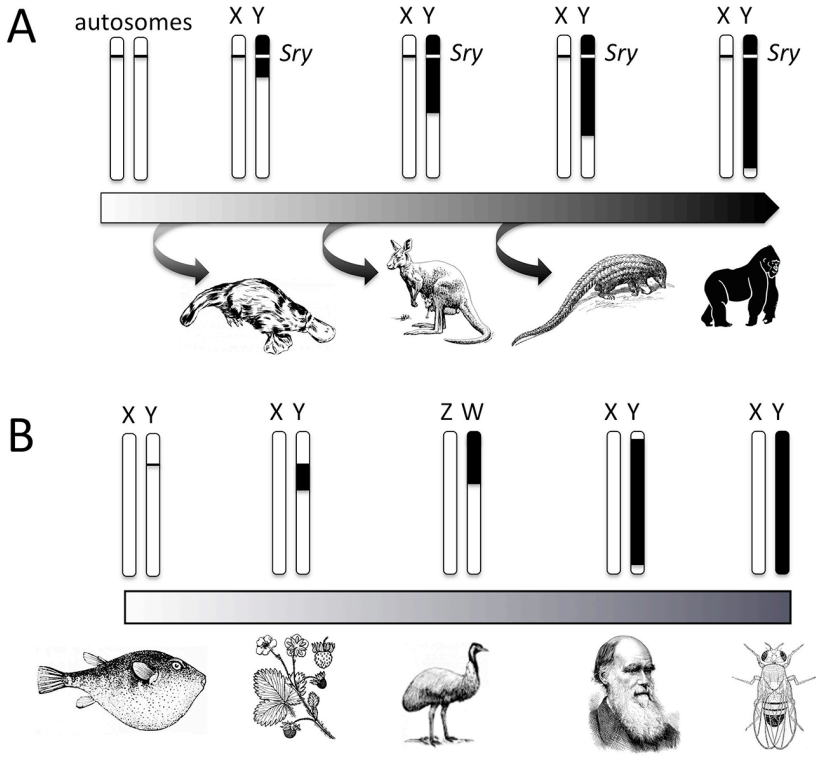

Figure 1. Illustration of the diversity and evolution of the sex chromosomes in some eukaryotes. A) Sex chromosomes of the XY SD system are proposed to start from a pair of autosomal like proto-sex chromosomes, followed by a gradual diversification including the emergence of the SD gene *Sry* during evolution. B) Illustration of difference and similarity of the divergence between heteromorphic sex chromosomes in different species. Black indicates the degenerated region of sex chromosome that lacks the ability to recombine, white denotes the autosome like region. Figure adapted from[16].

1.1.2 Sex determining genes

SD systems other than the ZW and XY systems exist. For instance, environmentally regulated SD systems are common in reptiles[17]. Many species have evolved more than two sex chromosomes or kept only one chromosome. Detailed reviews can be found

elsewhere[16,18,19]. The *Sry* gene on chromosome Y initiates testes development during embryogenesis in mammals[20,21]. Genes with cognate functions might also exist in other SD systems, for instance, *Dmrt1* on chromosome Z in chicken has been proposed as a candidate. Despite the fact that sex chromosomes among different SD systems show great variation, surprisingly, the molecular pathways that regulate the sex determination seem to have kept a high degree of conservation[22].

This indicates that molecular pathways are likely to be of more ancient origin than sex chromosomes themselves. What roles the SD genes have played during sex chromosome evolution has yet to be determined[23,24]. In order to approach this question, two difficulties need to be overcome: First, information on important stages of sex chromosome evolution that are required to recapture critical evolutionary events is missing. Second, degeneration of one sex chromosome makes the current sequencing output fragmented and less informative than the data derived from autosomes.

The benefits of adopting sex chromosomes remain to be systematically assessed on an evolutionary time scale[25,26]. In addition, the diversity and similarity between sex chromosomes within the same and between distinct SD systems are still puzzling biologists: What has triggered the diversification of SD systems and what forces have maintained the variations of SD systems? Do different systems have universal molecular features? Whether sex chromosomes in independently evolved SD systems have undergone similar evolutionary events? To answer the above questions, systematic studies need to be conducted.

1.1.3 Birds as a model to investigate sex chromosome evolution

Studies of the Y chromosome in mammals and the W chromosome in chicken have revealed many unique evolutionary features of sex chromosomes. The sex chromosomes are targeted by retrotransposons more often than autosomes. In addition, one of the sex chromosomes in the heterogametic sex frequently became shortened and largely inactive because of degeneration and transformation into constitutive heterochromatin. The degeneration of one chromosome caused by recombination repression (RS), in which chromosomes cease to recombine during prophase I, probably increased the genetic conflict between sex chromosomes through accumulating mutations beneficial for only one sex[27,28]. On the other hand, the ability to carry out crossover and exchange chromatin between heteromorphic sex chromosome pairs is limited within pseudo-autosomal regions (PARs) during synapsis. The remaining part of the sex chromosomes presents a stratified gradient of similarities with their

counterpart, e.g. Y and X, so called evolutionary strata[29,30]. The evolutionary strata between sex chromosomes seem to be a consequence of recombination suppression between the chromosome pair, and this feature is specific in sex chromosomes regardless of the type of SD systems. Again, the evolutionary force driving these features has not been identified, and some fundamental questions, such as whether the enrichment of repeats is a cause for, or the consequence of, recombination suppression remain unanswered.

Avian genomes are relatively stable and composed of fewer repetitive (10%-20%[31]) elements compared to the mammalian genomes (> 60% in the human genome[32]). The fewer translocations between chromosomes, also referred to as inter-chromosomal rearrangements, thus make it possible to recover and deduce evolutionary hallmarks. Birds existing today are commonly classified into two big groups: *Palaeognathae* and *Neognathae*, largely based on morphological differences. The latter group includes about 95% of the extant birds separated by short evolutionary distance, indicating a chance to uncover the evolutionary events affecting sex-chromosomes among moderately diverged species. Conversely, the former group has been placed at a basal position on the phylogenetic tree, and sex chromosomes of some birds, e.g. common ostrich and emu, do not possess a large portion of degeneration in one chromosome, as observed in chicken. Those exclusive features of avian genomes provide a diverse reservoir to investigate the evolution of sex chromosomes.

Except for the avian genomes, closely related species, such as cichlids in Lake Malawi, or pufferfish from the genus *Takifugu* that have undergone explosive speciation events and possibly still possess sex chromosomes without huge divergence, might provide various early stages of sex chromosomes, which can be used to characterize the consequences of shaping the SD genes' behavior in vertebrates[33–35]. The unique features and diversity of sex chromosomes make them good candidates to study the evolution of chromosomes, in addition to using distinct evolutionary trajectories to compare the autosomes between the two sexes.

1.1.4 Survey of the bird W chromosomes

Armed with powerful sequencing technologies, whole genome sequences of many species can be obtained in parallel within a relatively short time. Thus, by applying well-designed experiments and appropriate strategies, biologists are able to test evolutionary hypotheses and address fundamental evolutionary questions with comparative genomic toolkits[36]. While chicken has been historically used as a model organism to study avian genomes, the evolution of avian sex chromosomes remains largely unexplored because of highly degenerated W

chromosome[31]. Other birds have not been systematically described with respect to the unresolved phylogenetic relationships[37]. Hence, molecular features behind the morphological features of the sex chromosomes have not been systematically investigated among distinct lineages of the extant birds.

Although chromosomes Y and W have degenerated independently, studies are more frequently focused on the Y chromosome[38]. Whether the degeneration of W chromosome follows similar evolutionary trajectories as Y remains to be systematically investigated, and such a comprehensive characterization of avian sex chromosomes is important for the understanding of general sex chromosome evolution. Once the evolutionary patterns are characterized, the dynamics of gene loss can be measured and compared to that of the mammalian genomes (**Figure 2)**.

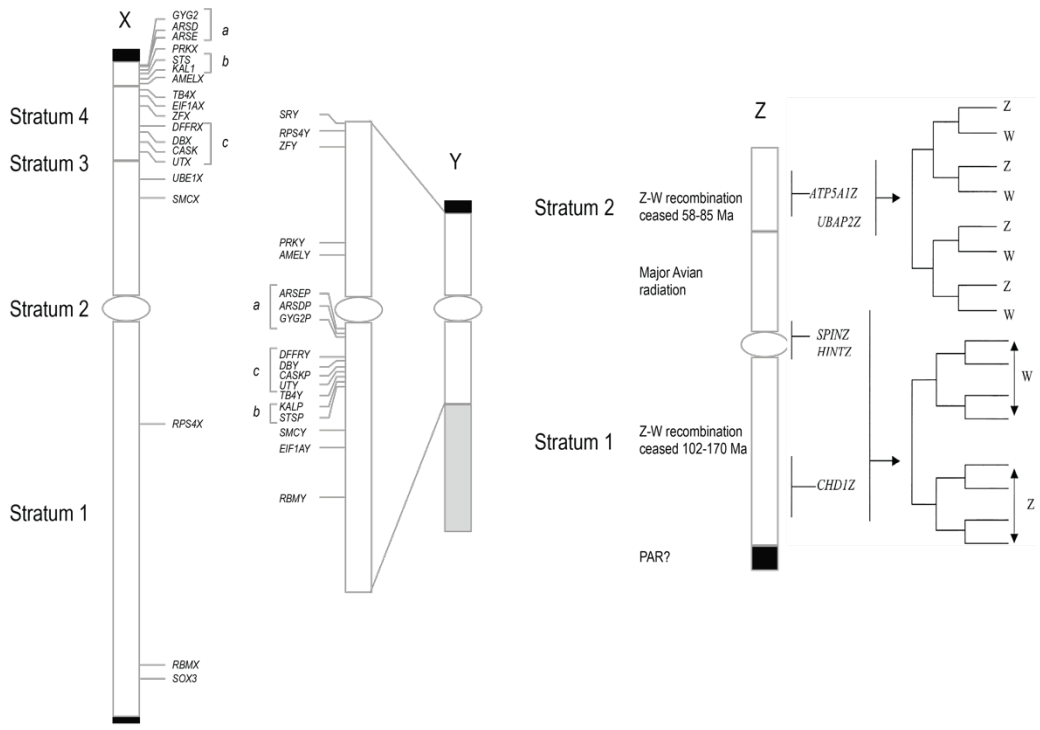

Figure 2. The convergent evolutionary strata of sex chromosomes in human and chicken genomes. Left. Evolutionary strata characterized on human chromosome X and corresponding genes on X as well as the genes on chromosome Y. The schematic centromeres are illustrated as white ovals and the heterochromatin is shown in grey. Gene clusters are indicated by brackets. PARs are coloured in black on both ends. Right. Evolutionary strata on chicken chromosome Z with the estimated divergence time. Corresponding genes with phylogenetic topologies present the distinct evolutionary histories

of genes between strata. Both Z and X present stratified patterns because of the stepwise decay of Y and W. Figure adapted from[29,30].

1.1.5 Strategy to identify the sex chromosome-linked sequences

Analogous to the Y chromosome, repetitive elements are highly enriched on W and the high degree of heterochromatinization of W makes it challenging to assemble the genomic sequence from short paired-end reads generated by the next-generation sequencing (NGS) technology. Despite these difficulties, it is still possible to recover a considerable proportion of W fragments by applying appropriate strategies. Theoretically, in a heteromorphic individual, the overall sequencing depth of each sex chromosome is expected to be half of the autosomal counterparts. Furthermore, considering the local bias introduced by the polymerase chain reaction (PCR) and incapability of recovering a long sequence for the repetitive regions, the reliability of approaches solely based on sequencing depth are sub-optimal[39,40].

Considering that sex chromosomes have evolved from a pair of identical autosomes, sequence alignment tools, such as LASTZ[41], can be tailored to identify sex-linked sequences in a genome with heteromorphic sex chromosomes by leveraging sequence homology. Together with the sequencing depth from a homomorphic individual, sex-linked sequences can be identified with high confidence.

1.1.6 Approach to improve the contiguity of genome assembly

Short read sequencing limits the contiguity and completeness of the sequences in the genome assembly. Optical mapping (OM) is one of the technologies used to organize and order the sequences derived from the same chromosome. The purified DNA is fragmented into super long DNA fragments followed by electrophoresis to place the single DNA molecules on a microfluidic device. After digestion by a carefully selected restriction enzyme, the digested DNA fragments are fluorescently labelled for microscopic detection to produce an enzyme map. Such maps derived from millions of single molecules are merged to generate an ultra-long enzyme map. By using dynamic programming to integrate *in silico* enzyme maps, the contiguity of genomic sequences can be significantly improved. For a detailed application, see[42]. Better contiguity is beneficial for deducting evolutionary histories by avoiding the ambiguities introduced by the fragmented sequences from the same chromosome.

1.1.7 Functional gene loss and non-coding regulators on one sex chromosome

The chicken W chromosome is highly degenerated, with less than 30 functional genes left[43]. Sequence analysis of the W genes and their counterparts on the Z chromosome (gametology) showed that the degeneration occurred following a stepwise pattern that formed evolutionary strata, which is similar to the Y chromosome[30].

It is still unclear why the fates of some genes on one sex chromosome were doomed during evolution while their counterparts survived[44]. A potential explanation might be a sexually antagonistic mechanism. In this scenario, the heteromorphic chromosome preserves genes that benefit one sex but do not benefit the other sex. However, whether this phenomenon is universal remains controversial, with one recent study suggesting that, at least in some species, this so-called sex antagonism (SA) is not the driving force of chromosome degeneration[45].

Besides the obvious importance of dissecting the functional roles of sex-linked genes, increasing evidence shows the importance of non-coding sequences. One of the known genomic loci in mammalian genomes that encodes the long non-coding RNA, *Xist*, plays an essential role in the inactivation of one X chromosome in females to balance gene dosage[46,47]. In avian genomes, a conserved male hyper-methylated (MHM) region in some lineages on the Z chromosome is postulated to be one of the non-coding regulators of sex determination during gonadogenesis[48,49].

Though recent sequence data from sex chromosomes have helped to elucidate evolutionary trajectories and gene dynamics, functional roles of the majority of the regulatory elements remain unidentified. Systematic characterization and functional investigation need to be performed to decipher their biological roles in order to address fundamental questions: Why are specific genes kept or silenced? What consequences did degeneration have? To what extent have the regulatory landscapes shifted from the original proto-sex chromosomes and have they contributed to the development of different SD systems? The genome sequence alone cannot address those intriguing questions without a comprehensive assignment and functional study of the genes, the regulatory elements (coding and non-coding regulatory loci), and the interactions between these elements and regulatory units (pathways).

1.2 POST-TRANSCRIPTIONAL REGULATION MEDIATED BY RNA BINDING PROTEINS

1.2.1 RNA binding proteins

The human genome encodes more than twenty thousand protein-coding genes, the majority of which are capable of generating various isoforms, either through alternative splicing or post-transcriptional modifications[50,51]. Although gene transcription is primarily controlled by transcription factors (TFs) that interact with *cis* elements through finely tuned temporal-spatial interactions, many other regulators modulate the consequences of primary transcription post-transcriptionally[52,53]. Given the fact that RNAs probably co-exist with those regulators longer compared to transcription, the fate of RNAs is possibly largely determined and guided by such regulators[54]. One group of proteins among those regulators, RNA binding proteins (RBPs), recognize and bind to RNA molecules. They participate in many essential post-transcriptional processes, suggesting their non-trivial roles in fine-tuning of RNA function[55] (see **Figure 3** for details).

Elucidating the roles of RBPs requires at least two key resources: 1) A catalogue of RBPs in the human genome, and 2) their binding specificities on the RNA sequences. The catalogue of RBPs keeps expanding, and the identification of their binding specificities remains challenging. As the majority of the genome can be transcribed into RNAs during development, and RBPs modulate the RNA metabolism, the poor understanding of RBPs hampers the decoding of functional elements embedded in the genome[56].

1.2.2 Major classes of RBPs

RBPs contain RNA binding domains (RBDs) can recognize and bind RNA through various mechanisms[57,58]. Although several types of RBDs have been described[59], recognition of RNA is generally accomplished by forming hydrogen bonds between RNA bases and/or RNA backbone with specific side chains of the folded protein, and the secondary structure or amino acids of RBDs is probably less relevant[60]. Despite the binding mechanism may not entirely relying on the RBDs, searching conserved protein domains is still a powerful approach to identify the RBPs in other species.

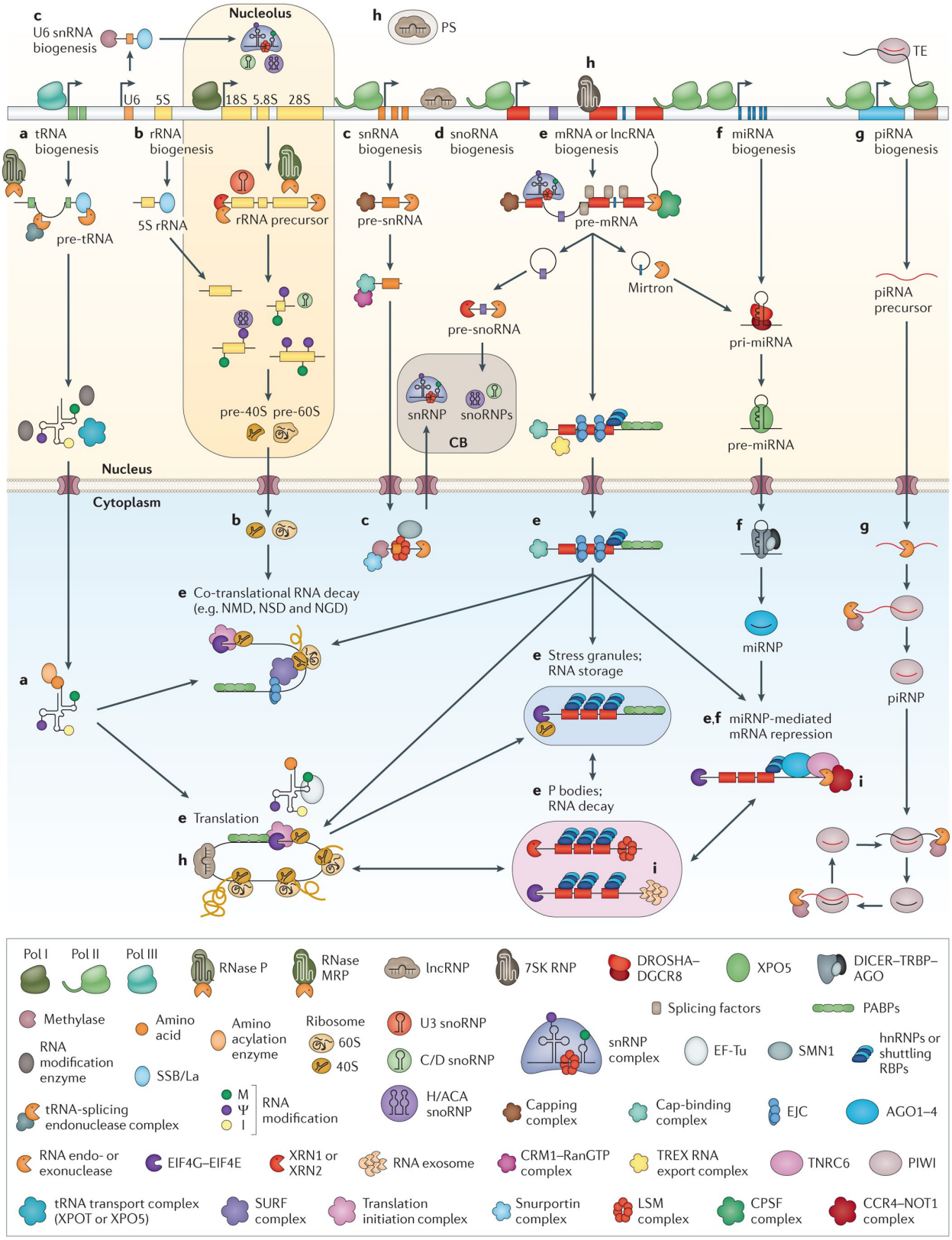

Figure 3. The complex roles of RNA binding proteins in post-transcriptional regulation. a, transfer RNA (tRNA) transcription; b, 5S ribosomal RNA (5s rRNA) transcription; c, small nuclear RNA (snRNA) transcription; d, small nucleolar (snoRNA) RNAs and small Cajal body-specific RNAs (scaRNAs); e, transcription of messenger RNAs; f, transcription of

microRNAs; g, Piwi-interacting (piRNA) transcription; h, transcription of most of the long non-coding RNAs; i, RNA degradation processes. Pol II/lll, RNA polymerase II/III; SMN1, survival motor neuron 1; CB, Cajal body; scaRNAs, small Cajal body-specific RNAs; miRNPs, miRNA-containing RNPs; PS, paraspeckles; EJC, exon junction complex; NGD, no-go decay; NMD, nonsense-mediatedRNA decay; NSD, non-stop decay; PABP, poly(A)-binding protein; TREX, transcription/export; XRN, exoribonuclease; LSM, like sm; SURF, SMG-1-Upf1-eRF1-eRF3 complex; AGO, argonaute. Figure adapted from[55].

RNA recognition motif

The RNA recognition motif (RRM), is the most frequently found RDB. It recognizes and binds to single-stranded RNAs, and is conserved in eukaryotes. The motif constitutes ~90 amino acids and folds into a βαββαβ structure with side chains interacting with the RNA[61–63]. The small size of RRMs makes them flexible building blocks and many RBPs contain more than one RRM copy, which increases diversity and flexibility to contact RNAs[64]. This modularity also allows the number of RRM domains to expand during evolution[65]. Hence, RBPs containing RRM domains usually combine various different binding specificities, which enables participation in multiple regulatory processes that determine the fate of RNAs, including transport, splicing, localization and translation[66–68].

K-homology domain

The K homology domain (KH) is a class of domain named after the heterogeneous nuclear ribonucleoprotein K (hnRNP K), where it was first described[69]. Subsequent proteins with this ~70 amino acid sequence were grouped into the KH domain family. RBPs incorporating the KH domain are frequently observed in metazoan genomes[58]. According to their secondary structure, KH domains can be further divided into two types. Type 1 contains a βααββα structure, whereas type 2 arranges the last two beta-sheets in the N-terminal instead, forming the αββααβ structure. Despite a short conserved motif GXXG located between helices 1 and 2 in type 1 and between helices 2 and 3 in type 2, those two types have only limited sequence similarity[70–72]. Similar to RRM domain containing RBPs, multiple KH domains are commonly observed, for instance in insulin like growth factor 2 mRNA binding protein 1 (IGF2BP1). Employment of several copies[58,70–72] or combinations with other RBDs, e.g. RRM, probably creates binding specificities that allow regulation of more specific processes[73] rather than less complex activities from single RBD[74,75].

Zinc fingers

Another frequent RBD type is the Zinc finger (Znf) domain, which is commonly included in relatively compact proteins with diverse binding domains, all of which either display DNA or RNA binding specificities with or without the requirement for metal ions[76]. The constitutional diversity of zinc finger domains among species and their non-constrained 3-dimensional structures indicate that Znf proteins may have arisen and evolved from distinct origins[77]. Binding specificities on RNAs of Znf proteins are generally difficult to determine, however examples where this has been achieved are ZFP36, MBLN and LIN28[78–82].

The diversity of binding domains enables a wide range of binding specificities. Recent studies have considerably expanded the catalogue of RBPs that lack canonical RBDs[55,83,84], however whether or not a certain experimental approach can distinguish true RNA binding proteins from RNA-associated proteins is under debate. The number of RBPs encoded in the human genome has increased to ~1500. Considering the complexity of RNA binding strategies and limitations of current experimental techniques, the number of RBPs has likely been underestimated[85], mainly due to their ill-defined properties[86–88].

1.2.3 Dysregulation of RBPs in post-transcriptional regulation

RBPs actively participate in RNA biogenesis and metabolism, including transcription, alternative splicing and RNA modification, transport, localization, translation and degradation[86–91]. Since RNAs localize both to the nucleus and the cytoplasm to exert their functions, it is necessary to precisely control the fate of RNA in order to avoid malfunction of subcellular components. Malfunction of RBPs has been implicated in neurodegenerative disorders, such as Alzheimer disease, frontotemporal dementia, and amyotrophic lateral sclerosis (ALS), as well as genetic diseases, e.g. Fragile X Syndrome and Myotonic dystrophy [92,93].

Dysregulation of splicing can have severe consequences, including malignancy[50,94]. A study utilizing a genome-wide CRISPR drop-out screen has demonstrated that elevated expression level of heterogeneous ribonucleoprotein HnRNPL facilitates prostate cancer progression[95]. Accurate splicing of the nascent RNAs into the correct isoforms needs tight coordination between spliceosomes and other RNA binding factors. RBFOX1 (RNA binding protein fox-1 homolog, also known as A2BP1/FOX1), plays for a central regulatory role in neuronal development by affecting downstream target transcription factors and synaptic proteins.

Rbfox1 itself alternatively splices into nuclear and cytoplasmic isoforms, and its cytoplasmic form helps to stabilize the target mRNAs in Autism[96,97].

One type of RNA modification is polyadenylation of messenger RNA transcripts, which occurs in eukaryotes by a two-step process. The cleavage and polyadenylation specificity factor (CPSF) protein first catalyses the cutting of precursor mRNA at a cleavage site between its conserved binding sequence AAUAAA and a downstream locus with degenerate U/GU -rich sequence. Then the poly(A) polymerase captures and processes the cleavage product by adding a poly(A) tail[98]. The endonuclease CSPF-73 is the main active component during cleavage event, and mutations of the key residues in its yeast homolog are lethal, which emphasizes the fact that RBPs are involved in highly conserved, fundamental functions[99]. Another prevalent and commonly studied modification is RNA editing, which results in the alteration from adenosine to inosine by a deamination process catalysed by adenosine deaminases acting on RNA (ADAR) proteins[100]. The base modification creates a more diverse mRNA population, which in turn also expands the proteome[101]. *In vivo*, RNA editing frequently occurs in the central nervous system, where many edited transcripts encode proteins functioning in rapid electrical and chemical neurotransmission that involves ion channels and G-protein coupled receptors. Decreased RNA editing of certain glutamate receptor transcripts in motor neurons has been observed in ALS patients, suggesting their strong association with the impaired function of adenosine deaminases acting on RNA 2 (ADAR2)[102].

RBPs also participate in the regulation of RNA localization[103,104]. The diversity of RNA binding domains possibly assists the precise control of RNA transport. A well-studied mRNA localization regulating protein zipcode-binding protein (ZBP1), which contains four KH domains, helps to move the beta-actin mRNA into lamellas regions by binding to the zipcode region at the 3' UTR of mRNA. Dysregulation of ZBP1 causes abnormal beta-actin mRNA localization in the cytoplasm[93,105].

In addition to the discussed roles of RBPs above, RBPs are key regulators of translation and mRNA turnover as well as many other biological processes. Overall, the broad scope of their regulatory roles and conserved functions imply that RBPs are an essential bridge to understand the causes of human disease, in particular, those caused by polymorphisms located in the transcribed non-coding sequences that could serve as binding sites to RBP (for example, 5' and 3' untranslated regions, lncRNA, etc).

1.2.4 *In vitro* approaches to determine RBP binding specificities

Numerous approaches have been designed and applied to deconvolute RNA-protein interactions, and, more importantly, to understand RBPs' physiological function by characterizing their binding specificities[106]. *In vivo* methods focus on the RBPs that bind to expressed RNAs, and can be generally grouped into two classes: cross-linking based protocols and cross-linking free protocols. In contrast, *in vitro* approaches can also identify binding specificities that have been wiped out from the genome due to negative selection. The increasing number of datasets generated from these types of experiments also enable computational methods to guide the discovery, even make precise predictions, of the RBP binding motifs.

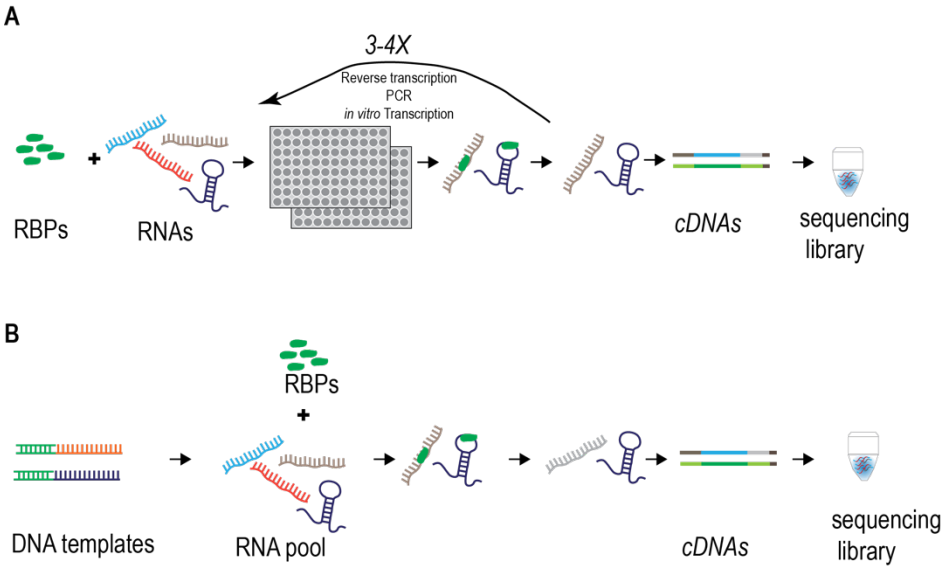

Figure 4. *In vitro* methods to determine the RBP binding specificities. A) Schematic illustration of the major high-throughput SELEX experiment; B) Design of the RNAcompete experiment. Both methods apply the concept of incubating proteins with RNA from large pools of RNA oligos. SELEX iterates the incubation procedure 3-4 times, whereas RNAcompete usually performs only one reaction.

Currently, the two most widely applied *in vitro* methods are sequencing-based assays, systematic evolution of ligands by exponential enrichment (SELEX) and the robust microarray-based assay RNAcompete[87,107–110]. Both strategies start from chemically synthesized DNA fragments to generate a RNA pool that is incubated with RBPs expressed

in *E. coli*, followed by purification and sequencing procedures (**Figure 4)**. Another approach, RNA Bind-n-Seq, is based on a similar general concept. However, RNA Bind-n-Seq has two additional interesting features: 1) Multiple RBP concentrations are considered in order to optimize within a range of affinity, and 2) the effects of RNA secondary structure on binding are assessed by a thermodynamically based approach[111].

All three methods can be used to determine the RBP-RNA interaction and to identify binding specificities, with the drawback of presenting motifs in non-physiological, *in vitro* conditions.

1.2.5 Motif discovery algorithm and representation of binding specificities

Computational approaches have also been developed to identify the binding sites of RBPs in RNA sequences, either with or without secondary structure information[112]. In addition to the tools designed for RBPs, motif discovery algorithms designed to identify TF binding motifs can be applied to search for RNA motifs, although the flexibility of RNA and complexity of the secondary structure poses challenges. One of these methods utilizes local maxima, defined by the top sub-sequences that have the highest counts in the sequence cluster within *Huddinge distance* equal to one. The *Huddinge distance H* can be written as $H=d-a$, where d denotes the maximum length of non-gapped bases and a represents the maximum length of properly aligned sequences. The sequences of local maxima are used to generate the initial binding profiles[113].

Position weight matrices (PWMs) are used to denote the fraction of nucleotide occurrences at each position in the motif. In a classic PWM matrix, row $R \in \{A, C, G, U/T\}$ and column $i \in \{1..N\}$, where N is the motif length. However, the PWM does not reflect dependency between nucleotides except for the fraction of nucleotides of independent positions in linear sequences.

1.2.6 Searching motif binding sites in the genome

Another main purpose of identifying the RBP motifs is to find binding sites in their target RNAs. Current computational tools, for instance, MOODs, CisGenome, MULTIPLESCAN and RBPmap[114–117] either use the consensus sequence or the PWM to search for the potential binding sites of TFs/RBPs. In principle, all these tools can detect potential binding sites if the motifs are linear sequences. Direct searching approaches for the sites with known structured motifs are not common mainly because an appropriate data presentation of the structured

motifs do not exist. Fast-algorithms, such as suffix tree or suffix array[118], can substantially accelerate the search for binding sites for RBPs binding to linear sequences. At the same time, the traditional online algorithm using the sequential subsequence strategy remains a straightforward approach for detecting structured binding sites of RBPs. More powerful computational tools will emerge once investigators have collected a high-confidence set of structured motifs.

Knowing the binding specificities of RBPs will assist the interpretation of the regulatory roles of both RBPs and corresponding RNA transcripts. Therefore, the interactions between functional elements and their contribution to the control of phenotypes can be further characterized and identified.

1.3 FUNCTIONAL ANNOTATION OF THE GENOME

The enhanced genome annotations and functional dissections in multiple dimensions have revolutionized our view of genomes[51,119,120]. Whole-genome sequencing projects have provided unique insights into human disease, food cultivation, and preservation of endangered species at a molecular level. In particular, sequencing of a huge number of individuals within certain populations to characterize mutations associated with disease, so-called whole genome associated studies (GWAS), has revealed that many human diseases are associated with genomic variations[121–124]. Moreover, RNA-seq enables the quantitative assessment of the associations between genomic loci and phenotypes through profiling of gene expressions[125].

Building the links between phenotypes and genomic loci is one of the top priorities for biologists. Many tools have been invented to investigate the functional roles of certain genomic loci at a local functional compartment or the whole genome level, for instance, the knockdown of gene expression approach with RNA interference (RNAi)[126], as well as the loss of function approaches mediated by engineering the genomic compositions: Zinc Fingers (ZNF)[127], transcription activator-like effector nucleases (TALEN)[128,129] and genome editing by CRISPR. The revolution of sequencing technology has greatly advanced our understanding of the genomic compositions by combining various omics data such as DNA, RNA and epigenetic modifications. Despite the expanding catalogue of functional elements is rapidly reshaping our view towards genomes[130], the code that controls the links between phenotypes and genotypes remains largely unknown and is challenging to deconvolute due to technical limitations. Existing methods *in vivo* in mouse such as site-specific recombinase technology and genome engineering can sort out the links, however, low throughput dramatically slows down progress. Within cultured cell systems, relatively simple experiments, such as genome wide knockout screens, make it easier to investigate the genes or elements that control phenotypes.

1.3.1 RNA interference

RNA interference (RNAi) is one of the most successful gene expression perturbation approaches, which utilizes the RNA induced silencing complex (RISC) to degrade mRNAs targeted by a short double strand RNA (dsRNA)[126,131]. It has been applied in various species to interrogate the functions of genes[132]. Although RNAi can interrogate the functional role of expressed genes, it is restricted to reducing the dose, but unable to create real null alleles. Off-target effects remain a concern and need to be appropriately assessed[133,134]. Despite these

shortcomings, RNAi still serves as a powerful tool to manipulate gene expression without permanently altering genomic content[135].

1.3.2 Genome editing

Until the emergence of CRISPR/Cas9, the lack of an efficient genome editing tool has greatly hindered the functional annotation of the genome. In the CRISPR/Cas9 system, one short 'guide' RNA is used to match a desired target DNA sequence in the genome, which also binds to the nuclease Cas9 that introduces a DNA double-strand break in the targeted DNA. This modified tool is derived from a naturally occurring genome editing system in bacteria, that has the ability to edit any genomic locus in any eukaryotic species. Compared to earlier genome editing tools, CRISPR/Cas9 and its derivatives are much faster and scalable[136,137]. In addition to protein-coding genes, several studies have successfully pioneered the precise function interrogation of non-coding regions with CRISPR-based tools and proved their feasibility to interrogate non-coding regulatory elements[7,138] (**Figure 5**). A review of the broad and extended application of CRISPR/Cas approaches can be found elsewhere[139].

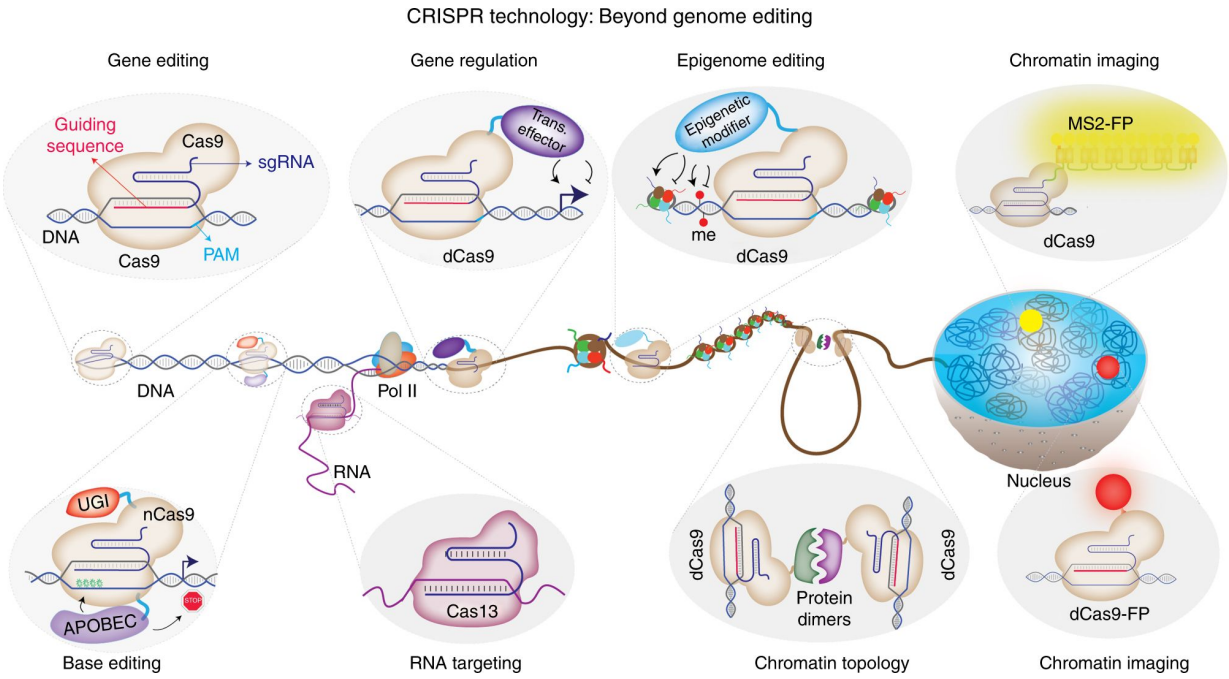

Figure 5. Derivatives of CRISPR/Cas9 beyond genome editing. Wild type Cas9 enables genome editing by cutting the DNA, while catalytically impaired Cas9 enzymes have been

applied to modulate gene regulation, for epigenome editing, chromatin imaging, and studying chromatin topology. The nickase Cas9 enzyme is used for base editing without introducing double strand breaks. The RNA targeting system is also implemented. Figure adapted from[140]. PAM, protospacer adjacent motif; UGI, uracil DNA glycosylase; FP, fluorescent proteins; APOBEC, APOBEC deaminase.

1.3.3 Precise genome-wide interrogation of gene function

Unlike the traditional mutagens that introduced mutations into the genome at random positions, CRISPR promises extremely high locus specificity by utilizing an exclusive guide-RNA to target the desired genomic locus with relatively low or undetectable off-target effects[141]. Investigators have leveraged the ease and high specificity of the CRISPR approach to implement a multiplexed screening method to perturb genes of interest, even the entire genome within one experiment[142,143].

A CRISPR/Cas9 based pooled screen usually starts from establishing a cell line with stable expression of the Cas9 protein. Then a library expressing guide RNAs is transduced lentivirally into cells at low multiplicity of infections. After subjecting the cells to a selection pressure, e.g, a cytotoxic compound, the surviving cells with integrated DNA cassette that encodes the guide sequence are collected to conduct the following experiment and amplified for the NGS sequencing. For a loss-of-function screen, the depletion of template guide sequences indicates their targeted genes are essential or required for cells to survive.

Pooled CRISPR/Cas9 loss-of-function screening has been successfully applied to identify the genes responsible for many phenotypes[5]. However, several fundamental questions remain to be addressed: Do cells interrogated with the same guide sequences exhibit exactly the same perturbation on phenotypes in the cell population? How do random drift and sub-sampling affect the identification of essential genes? To answer the above questions, a computational model needs to be able to measure the behaviours of individual cells.

1.3.4 Options of the statistical model to analyse screening data

To analyse screening data, and to call hit genes with high accuracy one has to choose an appropriate statistical model according to the experimental design. Many tools have been developed to optimize sensitivity while controlling false discovery. Three groups of statistical models have been applied to RNAi high-throughput screens: strictly standardized mean difference (SSMD), z-score and t-statistic. For a primary screen, calculation of SSMD with the median of absolute deviations (MAD) and median is more suitable where no replicate

exists in most of the scenarios. However, careful consideration of the statistical models before analysis is necessary to avoid inappropriate data interpretation[144].

1.3.5 Elucidate the physiological consequences of mutated regulatory elements

Despite many insights provided by cell models into the links between gene functions and phenotypes, it is difficult to directly assess the physiological consequences of the mutated sequence on the whole organism without performing *in vivo* perturbations. As most human disease-associated mutations are located outside protein-coding genes, assessing the physiological consequences in animal models might provide useful clues of non-coding elements in human genomes. For example, variations upstream of the oncogene *Myc* are associated with many types of human cancers[145–149]. The functional roles of these non-coding loci and their physiological roles remain to be systematically disclosed *in vivo*. In addition, the explosion of functional annotations through various genomic data allows for computational predictions. Several computational programs, for instance, ARVIN[150], FATHMM-MKL[151], GWAVA[152] and CADD[153], have been developed to unveil the functional roles of mutated non-coding elements in the human genome. Although predictions from the above tools can facilitate the functional annotation of the human genome, the expanding catalogue of human non-coding elements and increasing complexity of regulation network will demand a more robust and scalable computational approach to disentangle the roles of all regulatory elements in distinct environment context.

To summarize, the novel high-throughput sequencing techniques have empowered our ability to inspect the genome functions at multiple dimensions. We will understand the genome better by comparing the genomic features across evolution, deconvoluting the regulatory codes, and by mapping the relationship between genotypes and phenotypes. With these aims, further exploring the therapeutic strategies for human diseases, new tools and approaches with higher efficiency and accuracy eventually will be invented. Only through harnessing powerful tools, integrating multi-dimensional data and widening our visions, we can attempt to complete the comprehensive genotype-phenotype maps and solve regulatory rules of all genomic elements, even decode the entire set of rules encrypted in the book of life.

2 AIMS

The main goal of my work is to contribute to our understanding of how the genome works, more specifically to decode the rules of how the genomic sequence is used to determine the phenotypes. In my thesis projects, several distinct lines of work have been pursued, and each one has provided unique opportunities for me to approach interesting biological questions computationally. Although not directly related to each other, all of them are relevant to the overarching theme of dissecting genotype-phenotype relationships. The specific aims of each study are:

I. To characterize the evolutionary history of avian sex chromosomes.

II. To systematically identify the binding specificities of human RNA binding proteins.

III. To improve precision and accuracy for hit calling in pooled CRISPR/Cas9 screens.

IV. To investigate the *in vivo* functional roles of highly conserved regulatory non-coding elements in individuals.

3 METHODS

In this chapter, I briefly describe methods that I used to analyse the data and to implement the tools for each study. Protocols of the laboratory experiments are not included, however, they can be accessed together with detailed methods in the attached publications.

3.1 STUDY I

3.1.1 Improve the contiguity of ostrich genome with optical mapping

High molecular weight genomic DNA was extracted from a blood sample of a male ostrich in Kunming Zoo of China, and then passed to OpGen Inc. to collect a single molecule restriction map (SMRM). The SMRM and *in silico* restriction map were combined to link scaffolds into super-scaffolds, which substantially facilitated the scaffold assignment and orientation. A chromosome Z (chrZ) sequence was generated by orienting and connecting the optical mapping improved super-scaffolds based on the fluorescence *in situ* hybridization (FISH) results from previous work[154], with 600 Ns filled in between.

3.1.2 Pseudo-chromosome construction and identification of PARs

The detailed assembly and annotation of all the species are described in the comparative genomics study of all these species[155]. In brief, the raw reads of all species were assembled into scaffold sequences by SOAPdenovo[156], and gaps between contigs within scaffolds were filled in by GapCloser (http://sourceforge.net/projects/soapdenovo2/files/GapCloser/). The chicken genome (galGal13, http://www.genome.ucsc.edu) and its complete Z chromosome sequence[11] were taken as a reference to build the neognathae pseudo-chromosomal sequences and the ostrich genome from this study was used as a reference for other paleognaths. The position of gene DMRT1 was manually placed on the chicken Z chromosome according to its co-linearity with the neighbouring genes DMRT2 and DMRT3, which together show a conserved syntenic relationship within other vertebrate species[157,158]. In ostrich, the gene DMRT1 was placed and orientated according to the previous FISH result of ACO1[154], as these two genes located on the same super-scaffold. Repetitive elements in both reference genomes have been masked prior to further alignments using a consensus avian repeat library by RepeatMasker (http://repeatmasker.org). The whole genome alignments were constructed using LASTZ with parameter setting '--step=19 --hspthresh=2200 --inner=2000 --ydrop=3400 --gappedthresh=10000 --format=axt'[41] and a score matrix set for distant species comparison. Alignments were converted into a series of syntenic 'chains', 'net' and 'maf' results with different levels of alignment scores using UCSC Genome Browser's utilities (http://genomewiki.ucsc.edu/index.php/).

Based on the whole genome alignment, the scaffolds of query species were first ordered and placed according to their best aligned positions to the reference sequences, i.e. for each scaffold, at least 50% of the entire sequence was aligned in the LASTZ net results. The overall identity and coverage distributions were calculated for each scaffold along the reference sequence with a 10kb non-overlapped sliding window. With the distributions of coverage and identity from the aligned sequences, scaffolds within the lower 5% region of each distribution were removed to avoid spurious alignments. Finally, scaffolds were ordered and oriented into pseudo-chromosome sequences according to their unique positions on the reference.

The raw reads of each species were mapped to their pseudo-chromosome sequences by BWA[159] with the parameter set '-o 1 -e 50 -m 100000 -t 4 -i 15 -q 10 -I -k 0', adjusting for insert size of the Illumina sequencing library. The read depth was calculated using SAMtools[160] within each 100kb non-overlapping window and normalized against the median value of depths per single base pair throughout the entire genome, to allow comparison among species. The normalized depth was then converted by colorRampPalette in R and plotted against a 256-value colour gradient array ranging from 0 to 1 (blue to green) along the Z chromosome to display the PAR and differentiated region. Boundaries of PAR were determined by a significant shift of depth values between neighbouring windows.

3.1.3 Identification and validation of W-linked scaffolds

After excluding the best-aligned scaffolds used for building the Z chromosome, a second round of LASTZ alignment against the Z chromosome sequence was performed to identify candidate W-linked scaffolds, with a minimum length cutoff of 1kb, an alignment cutoff of at least 50% of the entire scaffold aligned, and 70% identity. An autosome (chr1) was also run with the same pipeline for comparison. To validate the sex linked sequences, the re-sequencing data were mapped to calculate the read depths of a male chicken (17 fold coverage)[161], a male Crested Ibis (40 fold sequencing coverage) and a male Emu[162] (10 fold coverage) onto their candidate W-linked sequences, chrZ and chr1 sequences with the same BWA parameter setting and SAMtools mentioned above. The distribution of normalized read depth of scaffolds between sexes and along individual genes to confirm the expected female specific pattern of W-linked scaffolds. A total of 98.9kb candidate chicken W-linked scaffolds were also aligned to the reference genome by Discontinuous MEGABLAST (http://www.ncbi.nlm.nih.gov/blast/html/megablast.html), only scaffolds with more than 50% of the sequence uniquely aligned at a minimum identity of 95% were assigned to certain chromosomes.

To enable a systematic comparison among species, Z-linked sequences of each species were scaled to the same length as the reference sequence of ostrich or chicken, by filling the

alignment gap with the same length of 'N's as the reference sequence based on the 'net' result of LASTZ. Then the pairwise alignment identity between the Z/W were calculated based on the 'maf' result and candidate W-linked scaffolds were orientated along the Z chromosome sequences. The read depth and alignment identity of each scaffold were visualized with color-codes. Moreover, the Z/W pairwise alignment identities along the Z chromosome with a non-overlapping 100kb window were calculated to determine the boundaries between the neighbouring strata when there was a significant difference of the identities or occurrences of W-linked scaffolds.

3.1.4 Identification and annotation of W-linked genes

The protein sequences of Z-linked genes were mapped to the W linked scaffolds with BLAT[163]. The best aligned (cutoff: identity>=70%, coverage>=50%) region with extended flanking sequences of 500 bp at both ends was then subjected to GeneWise[164] (-tfor -genesf -gff –sum) to search for an intact open reading frame (ORF). Gene models with disrupted ORF, where at least one premature stop codon or frame-shift mutation was introduced based on the GeneWise report, were defined as non-intact genes. The aligned CDS sequences of single-copy Z/W gametologous genes were generated by MUSCLE[165] and the poorly aligned regions were removed by Gblocks with an empirical parameter setting: -b4=5, -t=c, -e=-gb[166]. Here, only the alignments longer than 300bp were kept for subsequent phylogenetic tree construction using RAxML[167] to infer whether their residing evolutionary stratum is shared among species or specific to lineages. Transcript sequences were also predicted by GeneWise with a minimum length cutoff of 150bp. To compare the expressions of the sex-linked genes, RNA-seq reads of ostrich females' liver and brain tissues[168] were aligned to the transcript sequences using TopHat[169] and the expression levels were quantified by reads per kilo base per million (RPKM) on the basis of unique alignments.

The paired Z/W linked protein coding genes and their chicken orthologs were also aligned by MUSCLE[165] and poorly aligned blocks in the resulting alignments were filtered out by Gblocks[166]. To compare the evolutionary rates of Z linked and W linked genes, their ratio of nonsynonymous substitution rate vs. synonymous substitution rate was calculated by PAML.

3.1.5 Rearrangement analysis

To estimate the evolutionary history of rearrangement events on sex chromosomes, the gene synteny analyses among species were carried out to identify 1:1 reciprocal best orthologs between Green Anole Lizard[170], Boa Snake (Boa constrictor constrictor) (http://bioshare.bioinformatics.ucdavis.edu/Data/hcbxz0i7kg/Snake/)[171], and all the bird species produced in this study by BLASTP.

3.2 STUDY II

3.2.1 Sequencing and generation of motifs

To systematically characterize the binding specificities of human RNA binding proteins, a high-throughput RNA-SELEX (HTR-SELEX) was designed based on HT-SELEX[172]. The final PCR products from HTR-SELEX were prepared into sequencing libraries for sequencing on Illumina HiSeq 2000 (55 bp single reads). After de-multiplexing, the initial data were analysed with the Autoseed algorithm[113] that was further adapted to RNA analysis by taking into account only the transcribed strand and designating uracil rather than thymine. This method identified both gapped and ungapped *k-mers* that represent local maximal counts relative to similar sequences within their *Huddinge* neighborhood[113]. A preliminary motif was generated by using each such *k-mer* as a seed. This initial set of motifs is then refined manually to identify the final seeds to remove artefacts due to selection bottlenecks and common "aptamer" motifs that are enriched by the HTR-SELEX process itself, and motifs that are very similar to each other. We compared the recovered motifs to known motifs, to replicate experiments and experiments performed with paralogous proteins to evaluate the quality of initial motifs. Individual motifs that were not supported by replicate or prior experimental data were not included in the final dataset. Draft models were manually curated to identify successful experiments, and final models were generated using the refined seeds.

Autoseed detected more than one seed for many RBPs. For some RBPs, up to four seeds were used to generate a maximum of two unstructured and two structured motifs. The structured motif denotes the motif that forms a secondary structure whereas the unstructured motif denotes a linear RNA sequence. Of these, the motif with the largest number of seed matches using the multinomial setting was defined as the primary motif. The motif with the second largest number of matches was defined as the secondary motif. The counts of the motifs represent the prevalence of the corresponding motifs in the sequence pool. We only included primary and secondary motifs in subsequent analyses.

To search for RBPs that bind to dimeric motifs, the PWMs were visualized and examined to find direct repeat pattern of three or more base positions, with or without a gap between them. The presence of such repetitive patterns could be either due to dimeric binding, or the presence of two RBDs that bind to similar sequences in the same protein.

The correlation diagrams for each seed were visually investigated to find motifs that displayed a diagonal pattern, which were used to find structured motifs. For each structured motif, stem loop model (SLM) models were built from sequences matching the indicated seeds; a multinomial two setting was used to prevent the paired bases from influencing each other. Specifically, when the number of occurrences of each pair of bases was counted at the base-paired positions, neither of the paired bases was used to identify the sequences that were

analysed. To present the RNA motifs that form stem-loop (or hairpin) structures, we defined dinucleotide dependency matrices ($16 \times L$) to present the inter-correlation between base-pairing positions with probabilities. Each row presents the dinucleotide combination $R \in$ {AA, AC, AG, AU...UA, UC, UG, UU}, column $i \in \{1..L\}$, $L = N/2$ (N is even) or $L=1+N/2$ (N is odd). The SLMs were visualized either as the T-shaped logo or as a PWM type logo where the bases that constitute the stem were shaded based on the total fraction of A:U, G:C and G:U base pairs.

3.2.2 Motif mapping

To elucidate the function of the RBPs, all the motifs were mapped to the whole human genome (hg38). Different strategies were applied for the linear and the stem-loop motifs. For the linear motifs, motif matches were identified with MOODS[114] with the following parameter setting: --best-hits 300000 --no-snps. For the stem-loop motifs, a novel method was implemented to score sequences against the SLMs. The source code is available on GitHub: https://github.com/zhjilin/rmap.

We identified the 300,000 best scoring matches in the genome, and further included any matches that had the same score as the match with the lowest score, leading to at least 300,000 matches for each motif. The matches were then intersected with the annotated features from the ENSEMBL database (hg38, version 91), including the splicing donor, splicing acceptor the translation start codon, the translation stop codon and the transcription starting site. The above features were filtered in order to remove short introns (<50bp) to avoid the influence of adjacent splice sites, and features with non-intact or non-canonical start codon or stop codon. The filtered features were further extended 1kb both upstream and downstream in order to place the feature in the centre of all the intervals. The motif matches overlapping the features were counted using BEDTOOLS (version 2.15.0) and normalized by the total number of genomic matches for the corresponding motif.

3.2.3 Motif comparisons and GO analysis

The HTR-SELEX motifs were aligned to the publicly available datasets for comparison with a described method[173], which measures similarity between motifs calculated by SSTAT with default parameter setting. To determine whether RBPs with similar RBDs recognize and bind to similar targets, the sequences of the RBDs and their motifs were also compared. First, the RBPs were classified based on the type and number of RBDs. For each class, the amino-acid sequence of the RBPs starting from the first amino acid of the first RBD and ending at the last amino acid of the last RBD were extracted. The annotation of the RBDs were double confirmed by querying each amino acid sequence against the SMART database, and annotated the exact coordinates of the domains through the web-tools: http://smart.embl-heidelberg.de and http://smart.embl-heidelberg.de/smart/batch.pl. Sequence similarities and

trees were built using PRANK[174] (parameters: -d, -o, -showtree). The topology of the tree representing the similarity of the domain sequence was visualized using R (version 3.3.1).

The top 100 transcripts sorted by the score density of each RBP motif were extracted from the classes of transcripts that are enriched in motif matches for each RBP. These 100 transcripts were compared to the whole transcriptome to conduct the Gene Ontology (GO) enrichment analysis for each motif using the R package ClusterProfiler (version 3.0.5).

3.3 STUDY III

To enable lineage tracing in the pooled CRISPR/Cas9 screens that incorporated random sequence label (RSL), gene essentiality screens were performed. In these screens, a CRISPR guide library is transduced into a population of cells such that each cell on average contains one guide sequence-RSL combination. Cells were cultured for a total of 28 days after transduction, and 100 million cells were reseeded at each split. At least 50 million cells at day 4, day 14 and day 28 were collected to extract genomic DNA for the sequencing library preparation. Cells collected at day 4 were considered as the control time point. To analyse the above sequencing data and assess the performance of hit gene calling (i.e. genes that are essential for cell proliferation and/or survival), we developed a new analytical method.

3.3.1 Quality control and random sequence label (RSL) counting

To evaluate the behaviour of individual cells, the number of RSLs for each guide need to be counted. With the design of the sequencing library, the RSL and Sample ID appear at the reads Name field in the output fastq file. The RSL and guide sequences were first extracted from the reads Name and Sequence field respectively. The two sequences were subjected to filtering processes by meeting two criteria: 1) The guide sequence can be found in the original library design used for oligo synthesis, 2) No ambiguous base can be (present as N) in the RSL sequence. All the count tables were merged into one master table which contains the count of RSLs for each guide.

3.3.2 Implementation of the hit calling tool

To rank the genes, the data tables were processed in a control-experiment pairwise manner sequentially through the following steps:

Data normalization

Data were normalized to total read count: c_{ij} and t_{ij} represent the raw read counts for RSL-guide j in guide-set i for control (Day 4 after lentiviral transduction) and treatment (Day 28 after lentiviral transduction), respectively. The normalized read counts and are calculated by:

$$c'_{ij} = c_{ij} \frac{\sum_{ij}(c_{ij} + t_{ij})}{2 \sum_{ij} c_{ij}}$$

$$t'_{ij} = t_{ij} \frac{\sum_{ij}(c_{ij} + t_{ij})}{2 \sum_{ij} c_{ij}}$$

Median effect size and variability of the guide-sets

The effect size for each RSL-guide in guide-set *i* was defined as the log2 of the fold change between treatment count and control count. A pseudo-count 1 was added to the normalized count to handle total loss of an RSL-guide in the treatment sample:

$$ES_{ij} = log_2 \frac{t'_{ij} + 1}{c'_{ij} + 1}$$

Next, the median effect size for guide-set *i*, and the median of the absolute deviations (MAD) of all RSL-guides or bins *j* in guide-set *i* from MES_i were calculated. The factor 1.4826 was chosen such that the MAD is approximately equal to the standard deviation under the assumption of normal distribution[144].

$$MES_i = median\ ES_{ij}$$

$$MAD_i = 1.4826\ median_{\ j}\ |ES_{ij} - MES_i|$$

Median effect size and variability of the control guide-sets

Similarly, a single median effect size and MAD for the 101 non-targeting control guide-set were also calculated with the following formulas:

Median effect size of all non-targeting RSL-guides (MES_{CON}):

$$MES_{CON} = median_{ij} ES_{ij}^{NONT}$$

Median absolute deviation of all non-targeting RSL-guides (MAD_{CON}):

$$MAD_{CON} = 1.4826\ median_{\ ij}\ |ES_{ij}^{NONT} - MES_{CON}|$$

Strictly standardized mean difference and ranking score

Strictly standardized mean difference is a measure for the significance of the difference in behaviour of sample *i* and the non-targeting controls. It takes into account both the effect size and the variability of the data. For the non-targeting guides, the average score and standard deviation were also calculated for the hit calling.

$$SSMD_i = \frac{MES_i - MES_{CON}}{\sqrt{MAD_i^2 + MAD_{CON}^2}}$$

$$Score_i = MES_i |SSMD_i|$$

3.4 STUDY IV

To investigate the functional significance of regulatory enhancers upstream of *Myc* gene on 8q24, the cross-species conservation was assessed. The files of conservation score generated from 100 species were downloaded from UCSC (http://genome.ucsc.edu) for the human genome (version hg19) calculated based on the multiple species alignment though PhastCons. The overall conservation was calculated in a 500bp sliding window and the local conservation for certain loci in a 10bp window respectively.

4 RESULTS

4.1 STUDY I

To investigate sex chromosome evolution, I improved the contiguity of the ostrich genome with OM data to obtain a high quality reference chrZ sequence. The scaffold N50 increased from 3.5M in the NGS assembly to 17.5M in the new OM assembly. By combining the supper-scaffolds with the FISH markers from the previous work[154] a chromosomal level chrZ sequence for ostrich was reconstructed.

I developed a computational pipeline to search the sex linked sequence in the newly sequenced genomes[155] by combining sequence homology with sequencing depth. I identified both Z and W linked sequences for 16 species, including three *Paleognathaes* (Ostrich, Emu and White-tailed tinamou) and 13 birds from the sister *Neognathae* that includes *Galloanerses* (Chicken and Pekin duck) and *Neoaves* (Anna's hummingbird and Chimney swift).

To identify the regions on chromosome Z lacking recombination ability between Z and W (W degenerated), I used a sequencing depth-based approach to detect the recombining PAR along chromosome Z, where it retained the ability to recombine between chrZ and chrW. The divergence between chrZ and chrW is expected to be much higher than the similarities between the sequenced reads, thus the reads derived from PAR will yield an autosomal-like depth, while the W degenerated regions will show half of that. Indeed, we observed that the identified PAR along chrZ shows different lengths across species. More than 60% of the chrZ in Ostrich and Emu are PARs, but only ~1% of chrZ in white-tailed tinamou was retained for recombination. We confirmed the reliability of Ostrich PAR by comparing genes on the chrZ with the genes on the cytogenetic map. Contrary to most of the *neoaves* that have very short PARs, two species, killdeer (*Charadrius vociferus*) and white-tailed tropicbird (*Phaethon lepturus*), retained unexpectedly long PARs, suggesting that lineage specific re-arrangements have occurred independently around PAR in at least some birds (**Figure 6**).

Due to the repetitive nature of chrW, the W-linked sequences are usually assembled into short sequences. We systematically identified the W-linked sequences by filtering out sequences shorter than 1kb that mapped to chrZ. The identified W-linked sequences were further validated in three species by comparing the sequencing depth of the sex chromosomes between sexes. We annotated the W genes with a homologous gene searching strategy and obtained potential functional W genes.

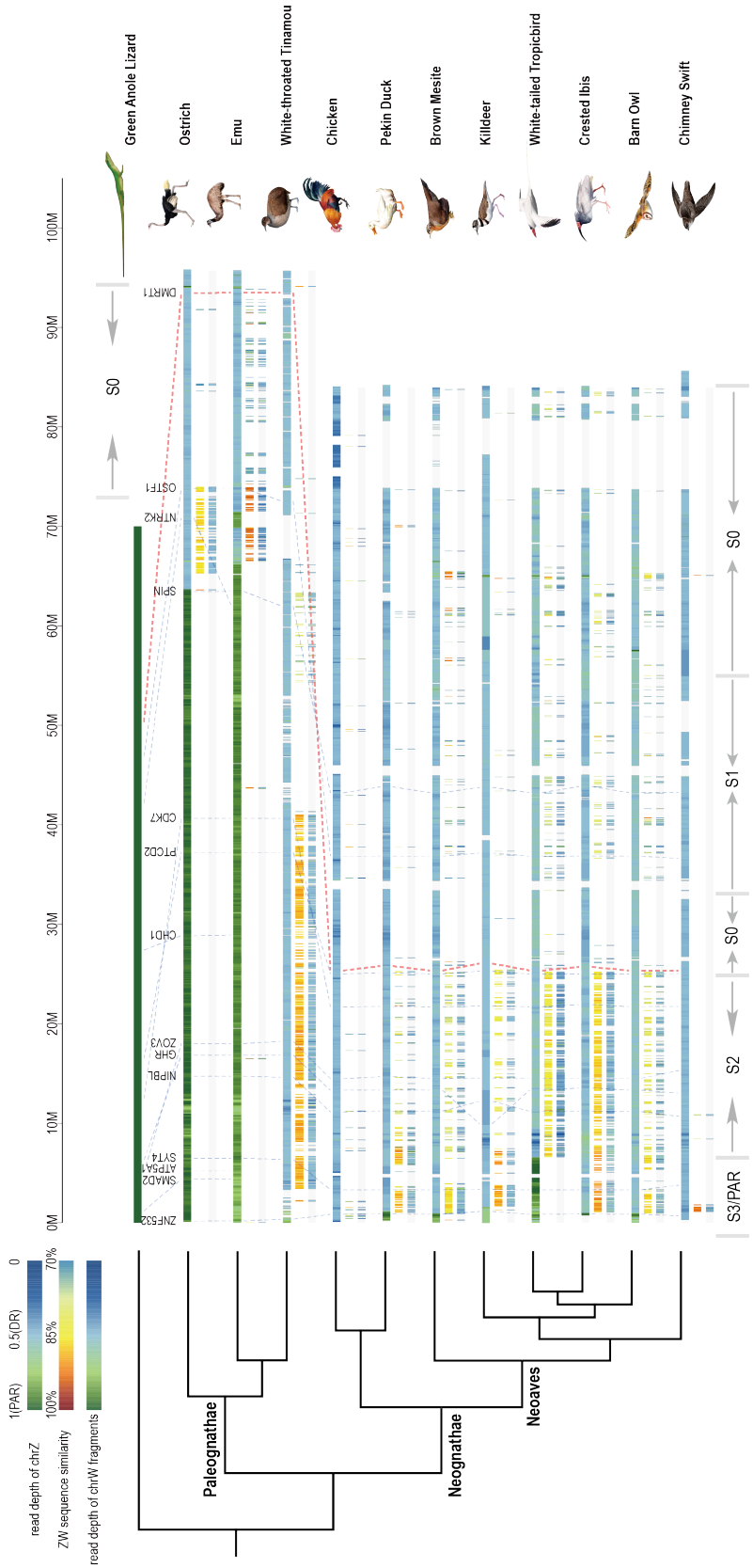

Figure 6. The evolutionary strata of avian sex chromosomes. The studied species are placed on a phylogenetic tree. The topmost track represents the normalized sequencing depth of chromosomes Z, the middle track represents the similarity between chrZ and chrW, and the bottom track represents normalized sequencing depth for each species. The regions of postulated strata are marked by two inverted arrows. Co-linearity of the genes in each stratum

is also indicated in each species and connected by the grey dotted lines. The SD candidate gene DMRT1 in avian genomes is marked by a red dotted line. Figure adapted from[175].

The W-linked sequences were placed along with their Z counterparts to assess the divergence between them. The divergence was scaled by the sequence similarity between homologous sequences. Consistent with the previously described evolutionary strata in chicken[30], the divergence between Z and W linked sequences is stratified into several segments according to the level of identity, which is most likely caused by recombination suppression events occurring at different times. By reordering the W linked sequence based on the ostrich chrZ (serving as a proto-Z), we approximately demarcated at least three strata, in which S0 and S3 represent the oldest and the latest stratum, respectively. Very few W-linked sequences were identified in S0, whereas more W-linked sequences were placed in the younger strata with considerably higher sequence similarities than that of the older strata. The older strata are most likely shared by all *Neoaves* species, suggesting the recombination occurred before the speciation events. Based on rescaled Z/W divergence level in noncoding regions, we estimated that stratum *Neognathae* S1 was approximately formed about 71 to 119 Ma, indicating it occurred before the split between *Neoaves* and *Galloanserae*.

In addition to the evidence obtained by sequence similarity, gametologous genes (homologous Z and W gene pairs) were also found to assist the recovery of evolutionary histories. If the recombination suppression occurred before the speciation, the Z and W gametologs are expected to be more diverged than gametologs that arose after the speciation. Therefore, gametologs within the common ancestral strata will group chrW genes and chrZ genes in two separate groups. Indeed, we identified gametologs that present such patterns within S0 and *Neoaves* S1 (**Figure 7)**.

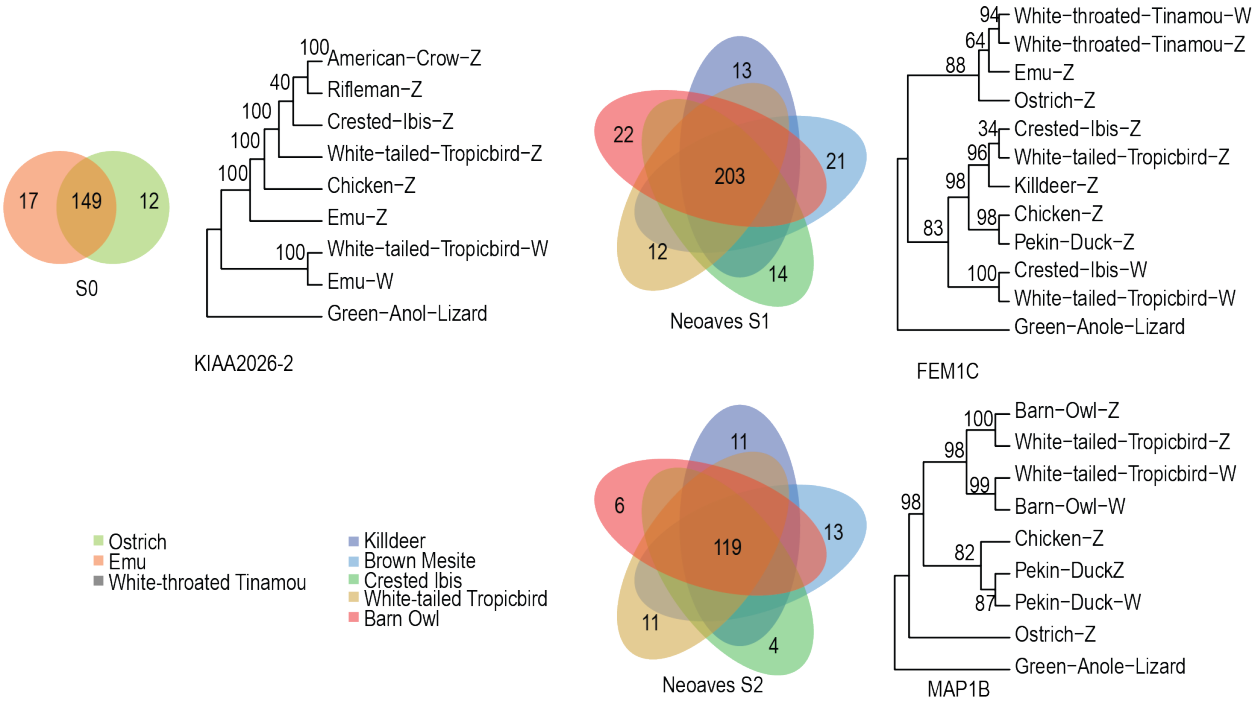

Figure 7. The emergence of the strata in distinct lineages. The Venn diagrams include the number of orthologous genes shared by the species within each stratum. A representative gene tree from each stratum is placed next to the Venn diagram to elucidate the evolutionary history of gametologs. Bootstrap values are placed next to corresponding nodes. Figure adapted from[175].

Further analysis of the gene synteny between avian genomes and comparison to the outgroup reptiles (green anole lizard and python) revealed that inversions on both chrZ and chrW might have contributed to the recombination suppression. A chrZ initiated inversion in S1 has relocated the putative sex-determining gene DMRT1 from the telomere proximal region into the middle of *Neoaves* chrZ close to the younger strata.

To conclude, I implemented a pipeline to characterize the diversity of sex chromosomes in avian genomes. The degree of chrW degeneration exhibits great variation among different lineages and at least 30% of the sampled *Neognathae* species harbour a chrW that is not as completely degenerated as that of chicken.

4.3 STUDY II

Binding specificities of the human RNA binding proteins

To systematically investigate the binding specificities of the human RBPs, **Arttu Jolma** collected 819 putative RBPs from Orfeome versions 3.1 and 8.1[176]. These contain canonical and non-canonical full length RBPs as well as RNA binding domains based on the annotation CisBP database[87] and RBPs list produced by Gerstberger et al.[55]. In brief, the RBPs were expressed as a fusion protein in *E.coli* and subjected to HTR-SELEX to identify the binding profiles[113,173] (**Figure 8A**).

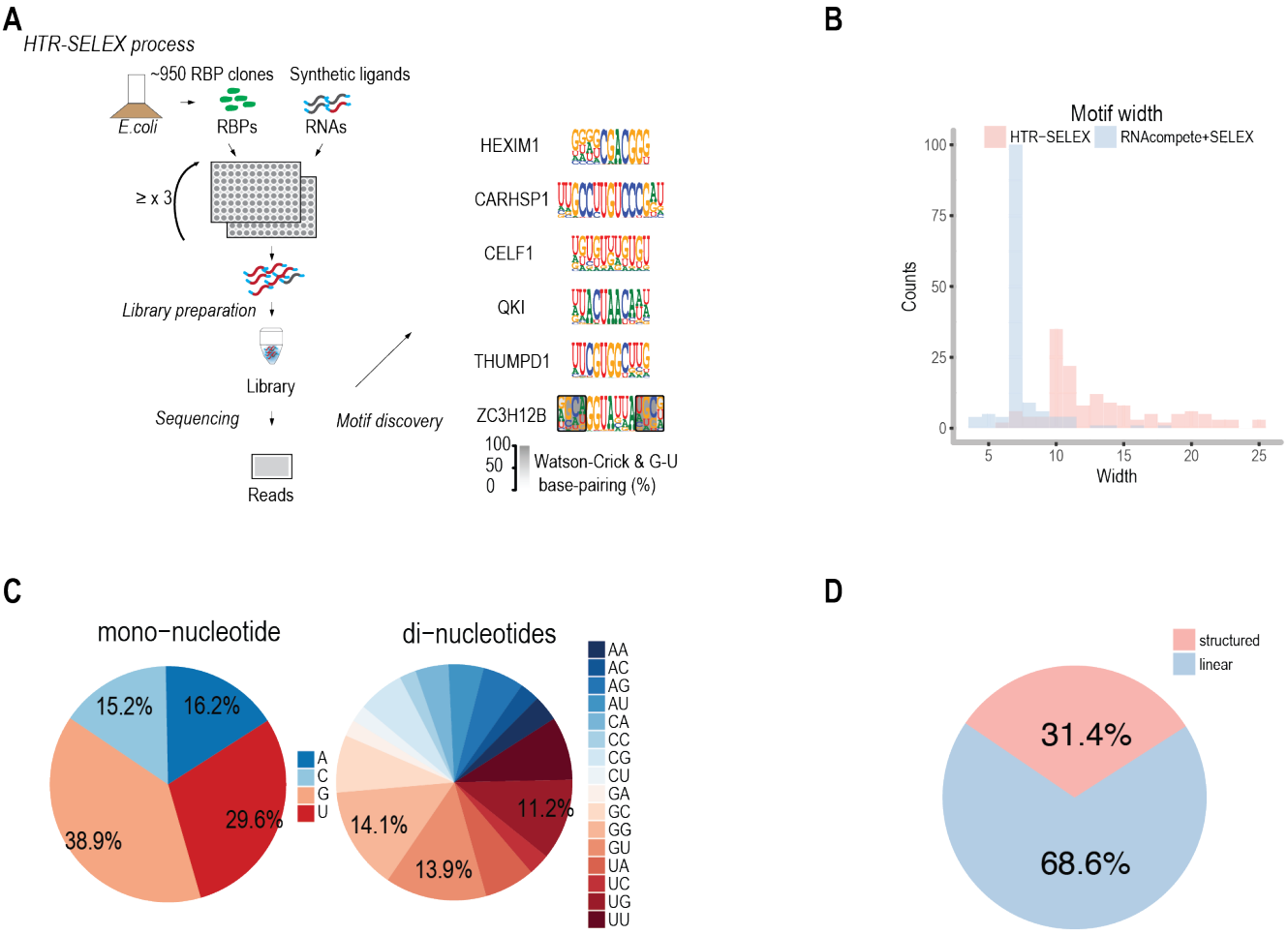

Figure 8. The HTR-SELEX and properties of the recovered motifs. A) Schematic illustration of the HTR-SELEX experimental procedure. The randomly synthesized DNA ligands are transcribed into RNAs through an *in vitro* transcription. The RNAs are incubated

with RBPs, followed by a washing process to remove the free RNAs. The RNAs oligos bound by RBP are disassociated and amplified to conduct a further incubation with RBPs. The whole procedure is repeated for at least three times before the sequencing library preparation. The sequencing reads are used to discover the motifs. B) The motif length distribution from HTR-SELEX (pink) and public available motifs (blue). C) The base composition bias in the HTR-SELEX motifs. D) The proportion of the linear and structured motifs.

To evaluate the strength and reliability of HTR-SELEX, I compared the identified 154 high-confidence distinct binding specificities from 86 RBPs with previously reported datasets, including RNAcompete[87], RNA bind-n-seq[177] and the data compiled in the RBPDB-database[178]. We found that more than 90% of the recovered motifs are from canonical RBPs, but also discovered motifs for 38 RBPs that had not been previously characterized. In addition, the motifs are generally consistent with previously reported motifs, although motifs generated by HTR-SELEX are much longer and have higher information content (**Figure 8B**).

Unexpectedly, despite displaying preferences towards structured RNA-sequences, about one-third of the RBPs are also able to bind the RNA sequences containing a direct repeat pattern. This suggests that some RBPs bind to RNA sequences through a cooperative strategy, for instance as homodimers, since these RBPs contain only a single RBD. A further survey of the distance between the direct repeats indicates that the distance between the RBDs is generally short (median 5 nucleotides), which is also observed in the *in vivo* data[179–181].

We further observed that the dinucleotides GG, GU, UG and UU in our motifs were more frequent than other dinucleotides (**Figure 8C**); fold change 2.75; $p < 0.00225$), suggesting that at least some RBPs prefer binding to G and U rich RNA sequences.

Structured binding specificities and motif characteristics

Based on the approach that generates dinucleotide dependency, we characterized the binding specificities of the 27 RBPs that can recognize structured RNA sequences[113] (**Figure 8D**). Analysis of the motifs revealed that 12 RBPs recognize both structured and linear RNA sequences, although most RBPs exclusively bind to either linear RNA sequences or sequences that form secondary structures. A closer examination indicated that RBPs from

RRM, CSD, Zinc finger and LA-domain families have the ability to bind both linear and structured sequences, whereas RBPs from the families KH and HEXIM bind to a linear sequence. To better model the stem-loop motifs, a stem-loop model (SLM) was designed, where loops are defined as a position independent model (PWM), whereas stems are treated as frequencies presenting each combination of two nucleotides from the paired positions. The information content defined by the SLM model was increased by 4.2 bits on average as compared to the linear model.

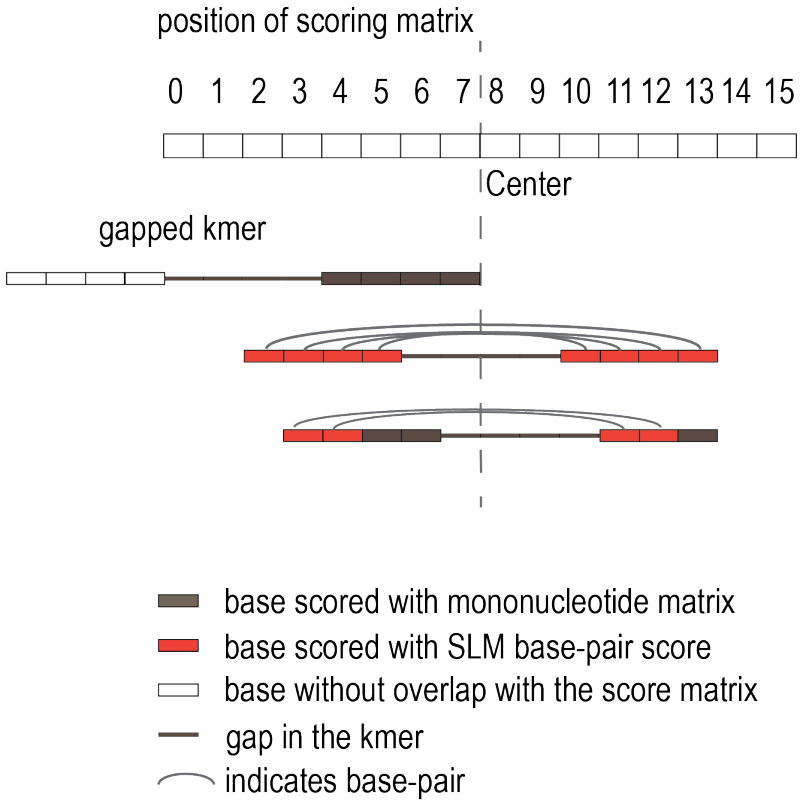

Figure 9. Schematic illustration of the mapping algorithm of structured motifs. The *k-mer* is aligned to the scoring matrix by searching for all the possible aligning positions in order to find the best-aligned position. The score of such alignment is used to represent the alignment between *k-mer* and the scoring matrix. For the structural model, the score is measured as the total scores of all aligned paired bases and unpaired bases. Normalization of the unaligned positions yields 0 in log odds, thus is not counted.

Revealing the functional roles of RBPs

To understand the functional roles of RBPs I searched the genome for potential binding sites of RBPs. Linear motifs can be mapped to the genome by MOODs[114], but a new online search method had to be established to map structured motifs (https://github.com/zhjilin/rmap)

(**Figure 9**). I selected the top ~300k genomic loci from the obtained motif matches for downstream enrichment analysis. The analysis of the normalized motif density around genomic features, including the transcription start site, splice donor and acceptor sites, and translational start and stop sites, revealed that many RBPs bind to the sequences around splice donor sites. Indeed, the binding sites of ZRANB2, one of the known splicing regulators, were greatly enriched around splice donors. Unexpectedly, we observed even stronger enrichment of structured motifs for ZC3H12A, B and C proteins around splice donor sites, suggesting their potential regulatory roles in splicing. Similar to the splice donor sites, several known RBPs involved in splicing were observed, such as RBM28, IGF2BP1 and ZFR[182–184].

In addition, we performed Gene Ontology Enrichment analysis to identify enriched binding sites in protein-coding genes. The motifs of many RBPs were enriched in transcripts exerting specific functions or close to splice junctions. For instance, MEX3B motifs were strongly enriched in genes involved in the type I interferon-mediated signalling pathway (**Figure 10**). Further conservation analysis of the motif matches was also conducted by **Teemu Kivioja** in mammalian genomic sequences close to splice junctions with the conservation score calculated by SiPhy[185]. The highly conserved binding sites of several RBPs in the transcripts indicates that those genomic loci containing the motifs are potentially under purifying selection.

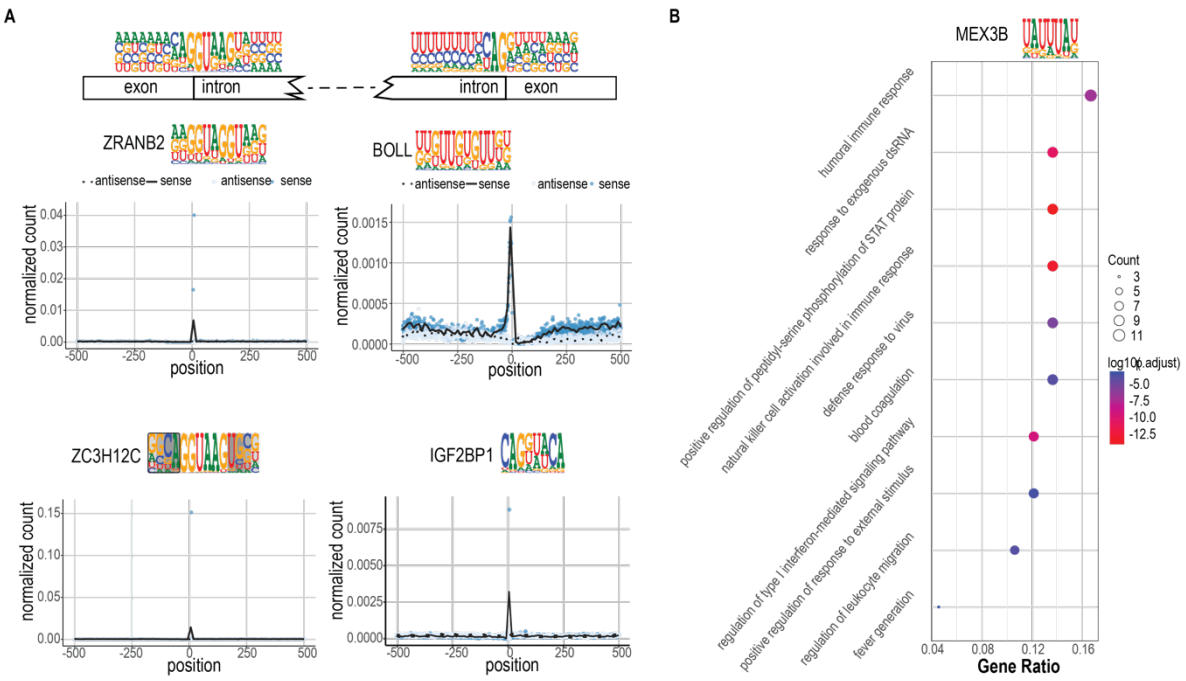

Figure 10. **Enrichment of RBP motif matches**. A) RBP motif matches are enriched at or close to splice junctions. The mononucleotide frequencies of splice donor and acceptor sites are shown on top of the schematic gene structure. The motif of each RBP is shown above

each subplot. Left: meta-plots indicate the enriched motif matches of ZRANB2 and ZC3H12C at splice donors. Right: motif matches enrichment of BOLL and IGF2BP1 at splice acceptor sites. The number of motif matches on the sense (blue) and anti-sense (light blue) strand at each base position is shown by dot; the locally weighted smoothing (LOESS) curves in 10 base sliding are shown as black line (sense) and dashed line (anti-sense). B) Enriched Gene Ontology terms of MEX3B motif matches. The top 100 genes with the highest score density were used to perform the GO enrichment analysis. Similar GO terms were merged according to their similarity with a threshold of 0.5.

To conclude, a collection of high-resolution motifs was obtained from the HTR-SELEX assay, of which many structured motifs have been identified. A novel motif mapping tool was developed to search potential RBP binding sites in the genome and bioinformatics analyses was further applied to explain the functional roles of RBPs.

4.4 STUDY III

Pooled CRISPR/Cas9 loss-of-function screens in cell models have been widely applied to simultaneously interrogate the effects of thousands of genes on a phenotype of interest. Noise due to differential behaviour of individual cells is a major problem in these screens, stemming from subsampling and random drift.

In this study, **Bernhard Schmierer** and **Sandeep Botla** incorporated random sequence labels (RSLs) to trace hundreds of individual virus-transduced cell lineages through a pooled CRISPR loss of function screen to elucidate genes essential for the human colorectal carcinoma cell line RKO. I implemented the analytical approaches (https://github.com/zhjilin/RSLC) to count the RSL-labelled guides. Afterwards, a median-based version of SSMD[186] was used to rank the guides to call the hit genes. To assess the performance of the method, we compared our results with the output of the commonly used tool MAGeCK[187] by feeding our data as the pooled screen data that lacks the RSL-labels. Our method outperformed the conventional approach, particularly when the cell number per guide was relatively low. Indeed, the analysis of the internal replicate hit calling revealed that the statistical power was greatly increased by including RSLs.

The experiments improved the hit gene calling by increasing the precision at least 15% on average compared to the internal replicate analysis, which is created through binning the RSLs. A much higher precision was observed at the lineage-specific level. To summarize, the RSL-guides have greatly increased the accuracy, precision as well as the statistical power in CRISPR/Cas9 based screening.

4.5 STUDY IV

In this study, **Kashyap Dave** and **Inderpreet Sur** conducted a series of large deletions in the regulatory region upstream of *Myc* gene through homologous recombination in mouse ES cells. A series of experiments were designed to study the functional roles of this region and found that enhancers within this region are not required for normal tissue development and possibly specific for the tumorigenesis.

I utilized the conservation score across distant species to indicate the functional importance of those enhancers has been implied by their sequence conservation. The ultra-conserved

features of these enhancer sequences suggest that they probably participate in ancient and thus fundamental functions (**Figure 11**).

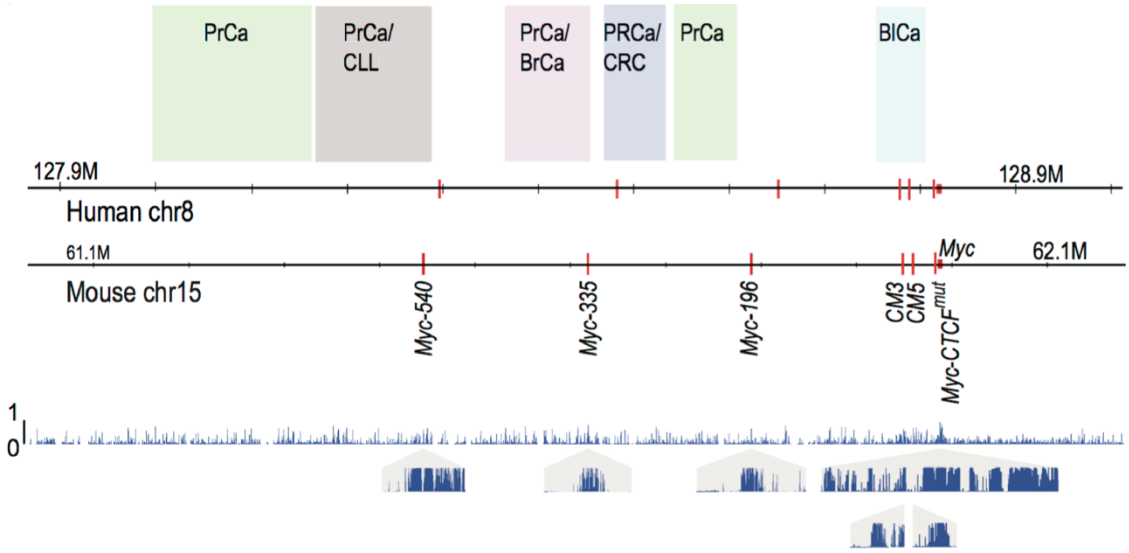

Figure 11. Conserved enhancer elements upstream of *Myc* that are associated with cancers. Top, the susceptibility regions for prostate cancer (PrCa), chronic lymphocytic leukemia (CLL), breast cancer (BrCa), colorectal cancer (CRC) and bladder cancer (BlCa) are marked. Middle, the comparison between human and mouse. The red vertical lines denote the locations of the Tcf7l2-binding CRC *Myc* enhancers. Bottom, the conservation probability of the entire region predicted by PhastCons (hg19 assembly, UCSC) with non-overlapping sliding windows and local conservation for each enhancer locus with a size of 500 bp and 10 bp, respectively. Figure adapted from[188].

5 DISCUSSION

5.1 STUDY I

In this study, I identified sex-linked sequences with a bioinformatics strategy based on sequence homology and sequencing depth. The results allowed us to demarcate true PAR regions on chrZ, and to identify a considerable length of diverged W-linked sequences. With the OM data and cytogenetic map, I generated a pseudo chromosome Z as the surrogate to chart and describe the evolutionary strata in other species. This method, however, discarded the local rearrangements on chrZ (e.g. small inversions) that might have contributed to the evolutionary histories of the sex chromosome. The complete sequence of Z chromosomes would advance the reconstruction of evolutionary events[11].

The total length of chrW sequence identified is underestimated due to limits imposed by the short reads from NGS and highly-repetitive features of sex chromosomes. However, this does not affect the identification of evolutionary strata, because neither local rearrangement on Z or incomplete W linked sequences could affect the trend of sequence similarity between Z and W that occurred in distinct evolutionary times.

I further constructed the phylogenetic trees of gametologous genes to assist the reconstruction of the evolutionary histories of the identified strata. Consistent with our observations, the genes on Z and W clustered based on homologous chromosomes rather than species in the ancestral stratum shared by all birds. Again, under-estimation of the W linked sequences resulted in fewer W linked genes, which lead to incomplete recovery of the evolutionary history. By the time of this study, anchoring the fragmented sequence without physical maps or linkage maps remains challenging and is difficult to obtain. With the current third generation sequencing and various long-range scaffolding techniques, complete sequence sex chromosomes will provide a better view of the evolution.

As in the mammalian sex chromosomes, the evolutionary strata are similar despite the Z-linkage of the potential SD gene DMRT1[11]. In addition, there is no evidence that the strata in avian genomes were induced by sex-autosome fusion as in mammals. Therefore, we propose that, if recombination suppression between chrZ and chrW was caused by large inversions, the events must have happened either on chrZ or chrW. This is in contrast to mammals, in which the recombination suppression might have happened only on chrY[189]. Although it has been proposed that the difference between systems might have been driven by the sex antagonism[190,191], it is still unclear whether this is universal. How the recombination

suppression has shaped gene function and their immediate consequences remain to be addressed with appropriate experiments.

To conclude, our analysis has provided a view of complex evolutionary trajectories in avian genomes. We found striking diversity of the sex chromosome composition among lineages. However, we do not know what has caused such diversities and whether we could explain the phenomenon with sex selection without further population data. Moreover, the molecular mechanism and consequences are yet to be explored to elucidate the underlying regulatory importance during embryogenesis.

5.2 STUDY II

We applied HTR-SELEX to determine the binding specificities of human RNA binding proteins, including both canonical and non-canonical RBPs[55,87]. The 40 nucleotide randomly synthesized RNA ligands allowed us to identify more complex binding motifs than previous studies with shorter ligands. We recovered stem-loop binding motifs for more than 30 RBPs. Most of the motifs were from canonical RBPs that recognize relatively short linear sequences. However, binding specificities from the HTR-SELEX are much longer with relatively higher information content[87,177,178]. Limited by the intrinsic properties of the experiments, we were not able to recover more complicated binding specificities, e.g. binding to sophisticated RNA secondary structures that require higher information content.

Half of the RBPs with motifs could recognize several different motifs, suggesting that the flexibility of a single RNA sequence probably enables a wider spectrum of binding specificity than DNA. More interestingly, the analysis of the motifs composition revealed an unusual bias towards G and U. The possible explanation might be the fact that U can also pair with G besides A in RNA sequences[192], thus decreasing their overall specificities.

Although most of the binding specificities we identified were linear, the successful identification of structured motifs suggests that RBPs recognize and bind to RNA via different strategies, which remains to be studied in detail. We noticed that the length of the RNA stem is generally short while the loop displays certain specificities, possibly because the relative short ligands cannot form complex structures. To our surprise, the RBPs are also able to bind RNA sequences cooperatively, similar to TFs on DNA[173]. However, the distance

between RBDs is generally short, suggesting that the flexible RNA sequences possibly require RBDs to cooperate in short distance to exert binding activity.

I explored the functional roles of RBPs via analysis of their potential binding sites in the genome and the conservation of these potential binding sites. The enrichment of binding sites around the splicing junctions suggested that some RBPs are very likely involved in the regulation of alternative splicing, as known splicing regulators as ZRANB2 display a similar pattern[193]. The enrichment of ZC3H12 proteins around the splicing junction suggested their potential anti-viral regulation in the cytosol mediated by CCCH-type zinc finger domain[194].

To summarize, we generated the largest collection of the binding specificities of human RBPs. Although the detailed functional roles of each RBP remain to be determined, we believe our results will provide a resourceful dataset for the community to facilitate the process.

5.3 STUDY III

The analytical framework I implemented for the hit gene calling in CRISPR/Cas9 screens with RSLs showed equal or better performance compared to the previous tools designed for pooled screens. As for any screening, only a subset of cells could be collected and assayed, some lineages are only present at one time point due to either subsampling or random drift. Besides the access of tracing lineage events, the labelling RSLs also makes it possible to control such effect by cleaning the data or building a probabilistic model.

To conclude, tethering RSLs into DNA templates for guide RNAs has greatly optimized the CRISPR/Cas9 screen. Even though we applied this in one knock-out screen to demonstrate its power, it has great potential to be incorporated with many other CRISPR/Cas9 screening methods and applications[7,138]. Unlike for genes, where several different guides can be used, the interrogation of non-coding genomic loci, where only one guide is available, relies on RSLs to generate replicates for the hit calling. Inclusion of RSLs does not require extra resources to achieve an equivalent analytical power, which is a substantial advantage in cases where screens are limited by scarce material (e.g. primary cells) or cost (large scale screens).

5.4 STUDY IV

To systematically investigate the role of region upstream *Myc* gene, several highly conserved enhancer elements in mouse were deleted, and two larger deletions were generated to

characterize the effects of deletions on both normal development and tumorigenesis. The close examination of sequence conservation across species has confirmed that those regulatory regions that hold functional significance are deeply conserved across vertebrates. However, the deletion of individual enhancers has little effect on MYC expression under the normal physiological conditions, which is consistent with the previous report[3].

6 CONCLUSIONS AND PERSPECTIVES

The four studies included here may seem to be only loosely related at first glance. Once step back and relook at their purposes, one may realize they are all subjected to address one question raised by Erwin Schrödinger: What is life? Indeed, the focus of each study is conducted at different magnitudes: molecular, subcellular, individual and cell population. All of them are answering questions at the different temporal spatial dimensions though different approaches and analytical strategies, which are essential building blocks to lay the road towards that ultimate BIG goal – to explain what life is.

We first quantitatively presented the diversity of sex chromosomes in bird genomes, characterized the evolutionary trajectories of sex chromosomes and proposed a model that describes the complex events occurred at different times to suppress the recombination between chrZ and chrW. Although we did not perform deeper analysis at the molecular level, our findings have provided a valuable example to use the genomes to address biological questions from an evolutionary perspective. The increasing volume of genomic data and rising of novel computational algorithms will bridge the gaps between our current observations and molecular mechanisms to understand genomes and their evolutionary histories.

To advance our understanding of the RBPs in the human genome, we applied the HTR-SELEX to identify their binding specificities. With this approach, we successfully characterized hundreds of distinct binding specificities for 86 RBPs. The analysis of those binding specificities disclosed important roles in many biological processes, including the regulation of alternative splicing, cytoplasmic antiviral defence as well as a potential universal binding mechanism between RBP and RNA. However, the functional roles of RBPs remain largely unexplored because of the limitation from either techniques or appropriate models to bind their targets. Emerging techniques, focusing on specific RNAs and their interacting proteins, will also provide more resourceful clues of how RBP exert functional roles in the specific context, which will further replete our understanding of the regulatory layer for the genome function.

The inclusion of UMI labelled single RNA guide has dramatically optimized the hit calling of CRISPR screens. Meanwhile, it also enabled the lineage dropout analysis and better control of the subsampling. Despite the current accumulated knowledge for dozens of genes that are related to human diseases, the function of the majority of the genes in our genome remain to

be unveiled. We believe that optimization of the CRISPR/Cas9 based screen will facilitate the *in vivo* dissection of functional roles of genes as well as other important non-coding genomic features.

To unwind the relationships between genotype and phenotype, functional characterization of non-coding regions is nontrivial. The *Myc* super-enhancer study not only elucidates the roles of enhancers in both normal and pathogenic conditions but also demonstrates an elegant example of designing appropriate experiments to perform the functional interrogation and validation for the non-coding genome. However, unlike the protein-coding sequences, to capture the functional roles of relatively less conserved non-coding sequences remains challenging, therefore, developing new tools and applying novel strategies are needed. Although CRISPR/Cas9 screens can provide resourceful functional interpretation, studies to test the physiological changes remain valuable and challenging. The fast-evolving genome editing tools probably will make the functional annotation much easier and scalable for a wide range of species besides model organisms, which substantially will advance our knowledge in regulation.

ACKNOWLEDGEMENTS

I would first like to thank my main supervisor and scientific mentor **Prof. Jussi Taipale**, who has provided me with such a unique opportunity to join his laboratory at Karolinska Institutet. I appreciate every discussion with you and your great patience to steer me back to the correct path, although you have limited time for the PhD students. Thank you for sharing with me your brilliant ideas and for providing me with all sorts of support during the past few years. I am very grateful to be a member of your laboratory to develop new skills. In a word, your wisdom and intelligence have enlightened me in many ways that keep motivating me to become a better independent researcher.

There are no words I could use to thank my co-supervisors **Drs. Minna Taipale** and **Bernhard Schmierer** for their enormous patience with me and the great help during my studies and research through my entire PhD journey. I appreciate that you have provided caring suggestions and advice both in my research and personal life.

I also want to thank all my kind collaborators in Taipale lab, **Drs. Inderpreet Kaur Sur**, **Eevi Kaasinen**, **Sandeep Botla**, **Ekaterina Morgunova**, **Arttu Jolma**, **Ning Wang**, **Yimeng Yin**, **Fangjie Zhu** and **Kashyap Dave**. I appreciate all the discussions of our research projects and great efforts to the papers. Without your contributions, I'm not able to do such wonderful research. From the bottom of my heart, you have given me a pleasant experience to work on the projects together with you. I hope we can continue the collaboration in the future.

I would also like to thank **Drs. Teemu Kivioja**, **Päivi Pihlajamaa** and **Biswajyoti Sahu,** my colleague and collaborator at Helsinki University**.** Thanks for your inspiring discussions with me, they helped a lot to solve my problems.

I also want to say a big thanks to **Drs. Lijuan Hu** and **Sandra Augsten** for the technical support I got the whole time. Thanks a million for taking care of my sequencing samples, and expressing proteins for the HT RNA-SELEX project.

Emma Inns, words are too pale to express my gratefulness. Thanks for all kinds of help you provided to me both in and outside the lab. Also, **Maria Hoh** and **Margareta Kling Pilström**, thanks for your great favours to help me to work smoothly with daily office and lab work.

Colleagues **Drs. Emma Haapaniemi**, **Kazuhiro Nitta**, **Åsa Kolterud**, **Jianping Liu**, **Fan Zhong**, **Otto Kauko**, **Alexander Minidis**, **Bei Wei**, **Jenna Persson** and **Anders Eriksson**. It was a really nice time to work with you in the same lab! I'm so grateful that you helped me when I turned to you. Thanks for the interesting discussions at the fika and lunchtime.

I would like to thank **Drs. Michael Lidschreiber**, **Katja Lidschreiber** and **Lisa Anna Jung** from Cramer Lab for sharing the adventures in Sweden besides the scientific discussions.

My gym buddies/friends**, Andreas Farde**, **Jacob Venuti Björkman**, **Carina Fischer**, **Tony Eklund**, **Per Nilsson**, **Mirela Ban** and **Chris Grigsby** my huge thanks to all of you, for bringing me into the climbing world besides safeguarding my life to the top (As it is supposed to be). My bouldering companions **Wing Leung**, **Daniela Mariosa**, **Albin Bohman**, **Jonathan Stenelöv**, **Ian Hoffecker**, **Giacomo Sitzia** and **Nora Björklund** for working together on the boulder problems. My lovely friends **Mohamed Ali**, **Catta Ramirez**, **Henrik Hedlund** and **Olmar El Mestikawy** who bridged and expanded my Swedish social circles.

My lovely corridor mates at Vårberg, **Sergio M. Martinez**, **Avril Madden**, **Robin Pronk**, **Ipsit Srivastava**, **Sara Albuquerque**, **Ana Pfitscher**, **Mandy Meijer**, **Marion Bodenmann**, **Ewoud Ewing**, **Rodrigo Marcondes** and **João Pedro Lopes** thanks for warming my heart with your kindness and sharing me the happy hours.

Thanks for all the support from my friends **Qian Zhou**, **Yimin He**, **Xing Li**, **Lihua Lu**, **Guangyi Fan**, **Zhe Ren**, **Sheng Jiang**, **Wenyan Li**, **Meng Ke**, **Xing Han** and **Feng Zhang**; and former colleagues or collaborators **Qi Zhou**, **Guojie Zhang** and **Changwei Shao** in China so I can keep moving during the dark time in my life.

I would like to thank all my friends **Hongya Han** and **Qiang Zhang** for helping me to settle down, and new friends **Xuan Li**, **Xinyan Miu**, **Yang Xuan**, **Jian Yan**, **Yin Lin**, **Janina Pfeil**, **Wenyi Zheng** and **David Van Bruggen** who brought me a great time at Karolinska Institutet.

Eva Nordlander, **Lina Rowland** and all other staff in the administration at bioNut and MBB, I very much appreciate your great help and patience with me for the past few years.

María García Collado, so much thanks to you for motivating, inspiring, supporting and understanding me during my development of the thesis.

My mentor, **Torbjörn Lundgren**, I appreciate the time you spent listening to me for the past two years when I was trapped in the valley of my life.

Lastly, I would like to thank my sister, parents and grandparents who have encouraged me to dive into the Scandinavian adventures**.** I also want to apologize for not being around the family for years. Without your love, support and encouragement, I would never walk this far by myself. I'm so grateful that you are always at my side, though with distance and time difference.

Finally, thanks to the grant agencies for the donation and grants. The entire PhD work was supported by grant KAW 2013.0088 from Knut and Alice Wallenberg Foundation to Prof. Jussi Taipale and KID funding led by Prof. Jussi Taipale.

Sincerely,

Jilin Zhang

Solna

September 2019

REFERENCES

1. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).

2. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).

3. Sur, I. K. *et al.* Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360–1363 (2012).

4. Li, L. & Chang, H. Y. Physiological roles of long noncoding RNAs: insight from knockout mice. *Trends Cell Biol.* **24**, 594–602 (2014).

5. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.* **16**, 299–311 (2015).

6. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).

7. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).

8. Sadleir, R. *The Reproduction of Vertebrates*. (Academic Press, 1973).

9. Speijer, D., Lukeš, J. & Eliáš, M. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8827–8834 (2015).

10. Fridolfsson, A. K. *et al.* Evolution of the avian sex chromosomes from an ancestral pair of autosomes. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 8147–8152 (1998).

11. Bellott, D. W. *et al.* Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466**, 612–616 (2010).

12. Nanda, I. *et al.* 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nat. Genet.* **21**, 258–259 (1999).

13. Cnaani, A. The tilapias' chromosomes influencing sex determination. *Cytogenetic and Genome Research* **141**, 195–205 (2013).

14. Jegalian, K. & Page, D. C. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394**, 776–780 (1998).

15. Graves, J. A. The origin and function of the mammalian Y chromosome and Y-borne genes--an evolving understanding. *Bioessays* **17**, 311–320 (1995).

16. Bachtrog, D. *et al.* Sex determination: why so many ways of doing it? *PLoS Biol.* **12**, e1001899 (2014).

17. Merchant-Larios, H. & Díaz-Hernández, V. Environmental sex determination mechanisms in reptiles. *Sex Dev.* **7**, 95–103 (2013).

18. Schultheis, C., Böhne, A., Schartl, M., Volff, J. N. & Galiana-Arnoux, D. Sex determination diversity and sex chromosome evolution in poeciliid fish. *Sex Dev.* **3**, 68–77 (2009).

19. Grützner, F. *et al.* In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. *Nature* **432**, 913–917 (2004).

20. Berta, P. *et al.* Genetic evidence equating SRY and the testis-determining factor. *Nature* **348**, 448–450 (1990).

21. Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. & Lovell-Badge, R. Male development of chromosomally female mice transgenic for Sry. *Nature* **351**, 117–121 (1991).

22. Raymond, C. S. *et al.* Evidence for evolutionary conservation of sex-determining genes. *Nature* **391**, 691–695 (1998).

23. Whitfield, L. S., Lovell-Badge, R. & Goodfellow, P. N. Rapid sequence evolution of the mammalian sex-determining gene SRY. *Nature* **364**, 713–715 (1993).

24. Nagai, K. Molecular evolution of Sry and Sox gene. *Gene* **270**, 161–169 (2001).

25. Kondrashov, A. S. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**, 435–440 (1988).

26. Agrawal, A. F. Sexual selection and the maintenance of sexual reproduction. *Nature* **411**, 692–695 (2001).

27. Helleu, Q. *et al.* Rapid evolution of a Y-chromosome heterochromatin protein underlies sex chromosome meiotic drive. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4110–4115 (2016).

28. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* **14**, 113–124 (2013).

29. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).

30. Handley, L.-J. L., Ceplitis, H. & Ellegren, H. Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution. *Genetics* **167**, 367–376 (2004).

31. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).

32. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).

33. Brawand, D. *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513**, 375–381 (2014).

34. Ieda, R. *et al.* Identification of the sex-determining locus in grass puffer (Takifugu niphobles) provides evidence for sex-chromosome turnover in a subset of Takifugu species. *PLoS One* **13**, e0190635 (2018).

35. Kamiya, T. *et al.* A Trans-Species Missense SNP in Amhr2 Is Associated with Sex Determination in the Tiger Pufferfish, Takifugu rubripes (Fugu). *PLoS Genetics* **8**, e1002798 (2012).

36. Hardison, R. C. Comparative genomics. *PLoS Biol.* **1**, E58 (2003).

37. Hackett, S. J. *et al.* A phylogenomic study of birds reveals their evolutionary history. *Science* **320**, 1763–1768 (2008).

38. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).

39. Chen, N., Bellott, D. W., Page, D. C. & Clark, A. G. Identification of avian W-linked contigs by short-read sequencing. *BMC Genomics* **13**, 183 (2012).

40. Carvalho, A. B. & Clark, A. G. Efficient identification of Y chromosome sequences in the human and Drosophila genomes. *Genome Res.* **23**, 1894–1907 (2013).

41. Harris, R. S. Improved Pairwise Alignment of Genomic DNA. (etda.libraries.psu.edu, 2007).

42. Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat. Biotechnol.* **31**, 135–141 (2013).

43. Wright, A. E., Moghadam, H. K. & Mank, J. E. Trade-off between selection for dosage compensation and masculinization on the avian Z chromosome. *Genetics* **192**, 1433–1445 (2012).

44. Hughes, J. F. *et al.* Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82–86 (2012).

45. Bergero, R., Gardner, J., Bader, B., Yong, L. & Charlesworth, D. Exaggerated heterochiasmy in a fish with sex-linked male coloration polymorphisms. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6924–6931 (2019).

46. Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A. & Panning, B. Xist RNA and the mechanism of X chromosome inactivation. *Annu. Rev. Genet.* **36**, 233–278 (2002).

47. Panning, B., Dausman, J. & Jaenisch, R. X chromosome inactivation is mediated by Xist RNA stabilization. *Cell* **90**, 907–916 (1997).

48. Yang, X. *et al.* A Window of MHM Demethylation Correlates with Key Events in Gonadal Differentiation in the Chicken. *Sex Dev.* **10**, 152–158 (2016).

49. Wright, A. E., Zimmer, F., Harrison, P. W. & Mank, J. E. Conservation of Regional Variation in Sex-Specific Sex Chromosome Regulation. *Genetics* **201**, 587–598 (2015).

50. Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72**, 291–336 (2003).

51. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

52. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* **13**, 233–245 (2012).

53. Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature* **468**, 911–920 (2010).

54. Geisberg, J. V., Moqtaderi, Z., Fan, X., Ozsolak, F. & Struhl, K. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* **156**, 812–824 (2014).

55. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).

56. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

57. Jankowsky, E. & Harris, M. E. Specificity and nonspecificity in RNA-protein interactions. *Nat. Rev. Mol. Cell Biol.* **16**, 533–544 (2015).

58. Burd, C. G. & Dreyfuss, G. RNA binding specificity of hnRNP A1: significance of hnRNP A1 high-affinity binding sites in pre-mRNA splicing. *EMBO J.* **13**, 1197–1204 (1994).

59. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**, 479–490 (2007).

60. Auweter, S. D., Oberstrass, F. C. & Allain, F. H.-T. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res.* **34**, 4943–4959 (2006).

61. Chambers, J. C., Kenan, D., Martin, B. J. & Keene, J. D. Genomic structure and amino acid sequence domains of the human La autoantigen. *J. Biol. Chem.* **263**, 18043–18051 (1988).

62. Dreyfuss, G., Swanson, M. S. & Piñol-Roma, S. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem. Sci.* **13**, 86–91 (1988).

63. Sachs, A. B., Davis, R. W. & Kornberg, R. D. A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. *Molecular and Cellular Biology* **7**, 3268–3276 (1987).

64. Deo, R. C., Bonanno, J. B., Sonenberg, N. & Burley, S. K. Recognition of Polyadenylate RNA by the Poly(A)-Binding Protein. *Cell* **98**, 835–845 (1999).

65. Anantharaman, V., Koonin, E. V. & Aravind, L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.* **30**, 1427–1464 (2002).

66. Sawicka, K., Bushell, M., Spriggs, K. A. & Willis, A. E. Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem. Soc. Trans.* **36**, 641–647 (2008).

67. Simone, L. E. & Keene, J. D. Mechanisms coordinating ELAV/Hu mRNA regulons. *Curr. Opin. Genet. Dev.* **23**, 35–43 (2013).

68. Query, C. C., Bentley, R. C. & Keene, J. D. A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. *Cell* **57**, 89–101 (1989).

69. Siomi, H., Matunis, M. J., Michael, W. M. & Dreyfuss, G. The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. *Nucleic Acids Res.* **21**, 1193–1198 (1993).

70. Baber, J. L., Libutti, D., Levens, D. & Tjandra, N. High precision solution structure of the C-terminal KH domain of heterogeneous nuclear ribonucleoprotein K, a c-myc transcription factor. *J. Mol. Biol.* **289**, 949–962 (1999).

71. Grishin, N. V. KH domain: one motif, two folds. *Nucleic Acids Res.* **29**, 638–643 (2001).

72. Musco, G. *et al.* The solution structure of the first KH domain of FMR1, the protein responsible for the fragile X syndrome. *Nature Structural Biology* **4**, 712–716 (1997).

73. Bell, J. L. *et al.* Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell. Mol. Life Sci.* **70**, 2657–2675 (2013).

74. Lasko, P. Gene regulation at the RNA layer: RNA binding proteins in intercellular signaling networks. *Sci. STKE* **2003**, RE6 (2003).

75. Chénard, C. A. & Richard, S. New implications for the QUAKING RNA binding protein in human disease. *J. Neurosci. Res.* **86**, 233–242 (2008).

76. Berg, J. M. Zinc fingers and other metal-binding domains. *Receptor* **29**, 31 (1990).

77. Emerson, R. O. & Thomas, J. H. Adaptive evolution in zinc finger transcription factors. *PLoS Genet.* **5**, e1000325 (2009).

78. Sandler, H. & Stoecklin, G. Control of mRNA decay by phosphorylation of tristetraprolin. *Biochem. Soc. Trans.* **36**, 491–496 (2008).

79. Brooks, S. A. & Blackshear, P. J. Tristetraprolin (TTP): interactions with mRNA and proteins, and current thoughts on mechanisms of action. *Biochim. Biophys. Acta* **1829**, 666–679 (2013).

80. Thornton, J. E. & Gregory, R. I. How does Lin28 let-7 control development and disease? *Trends Cell Biol.* **22**, 474–482 (2012).

81. Wilbert, M. L. *et al.* LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. *Mol. Cell* **48**, 195–206 (2012).

82. Cho, J. *et al.* LIN28A is a suppressor of ER-associated translation in embryonic stem cells. *Cell* **151**, 765–777 (2012).

83. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).

84. Ghosh, P. & Sowdhamini, R. Genome-wide survey of putative RNA-binding proteins encoded in the human proteome. *Mol. Biosyst.* **12**, 532–540 (2016).

85. Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1122–1130 (2013).

86. Zheng, D. & Tian, B. RNA-binding proteins in regulation of alternative cleavage and polyadenylation. *Adv. Exp. Med. Biol.* **825**, 97–127 (2014).

87. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).

88. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986 (2008).

89. Doyle, F. & Tenenbaum, S. A. Trans-regulation of RNA-binding protein motifs by microRNA. *Front. Genet.* **5**, 79 (2014).

90. Hasan, A., Cotobal, C., Duncan, C. D. S. & Mata, J. Systematic analysis of the role of RNA-binding proteins in the regulation of RNA stability. *PLoS Genet.* **10**, e1004684 (2014).

91. Chabot, B. & Shkreta, L. Defective control of pre–messenger RNA splicing in human disease. *J. Cell Biol.* (2016).

92. Svetoni, F., Frisone, P. & Paronetto, M. P. Role of FET proteins in neurodegenerative

disorders. *RNA Biol.* **13**, 1089–1102 (2016).

93. Wang, E. T. *et al.* Dysregulation of mRNA Localization and Translation in Genetic Disease. *J. Neurosci.* **36**, 11418–11426 (2016).

94. Sebestyén, E. *et al.* Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res.* **26**, 732–744 (2016).

95. Fei, T. *et al.* Genome-wide CRISPR screen identifies HNRNPL as a prostate cancer dependency regulating RNA splicing. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E5207–E5215 (2017).

96. Hamada, N. *et al.* Essential role of the nuclear isoform of RBFOX1, a candidate gene for autism spectrum disorders, in the brain development. *Sci. Rep.* **6**, 30805 (2016).

97. Lee, J.-A. *et al.* Cytoplasmic Rbfox1 Regulates the Expression of Synaptic and Autism-Related Genes. *Neuron* **89**, 113–128 (2016).

98. Minvielle-Sebastia, L. & Keller, W. mRNA polyadenylation and its coupling to other RNA processing reactions and to transcription. *Curr. Opin. Cell Biol.* **11**, 352–357 (1999).

99. Ryan, K. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. *RNA* **10**, 565–573 (2004).

100. Hoopengardner, B., Bhalla, T., Staber, C. & Reenan, R. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301**, 832–836 (2003).

101. Modrek, B. & Lee, C. J. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* **34**, 177–180 (2003).

102. Hideyama, T. *et al.* Induced loss of ADAR2 engenders slow death of motor neurons from Q/R site-unedited GluR2. *J. Neurosci.* **30**, 11917–11925 (2010).

103. Besse, F. & Ephrussi, A. Translational control of localized mRNAs: restricting protein synthesis in space and time. *Nat. Rev. Mol. Cell Biol.* **9**, 971–980 (2008).

104. Meignin, C. & Davis, I. UAP56 RNA helicase is required for axis specification and cytoplasmic mRNA localization in Drosophila. *Developmental Biology* **315**, 89–98 (2008).

105. Oleynikov, Y. & Singer, R. H. Real-Time Visualization of ZBP1 Association with β-Actin mRNA during Transcription and Localization. *Curr. Biol.* **13**, 199–207 (2003).

106. Riley, K. J. & Steitz, J. A. The 'Observer Effect' in Genome-wide Surveys of Protein-RNA Interactions. *Mol. Cell* **49**, 601–604 (2013).

107. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).

108. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).

109. Cook, K. B. *et al.* RNAcompete-S: Combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step in vitro selection. *Methods* **126**, 18–28 (2017).

110. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20**, 861–873 (2010).

111. Lambert, N. *et al.* RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol. Cell* **54**, 887–900 (2014).

112. Li, X., Kazan, H., Lipshitz, H. D. & Morris, Q. D. Finding the target sites of RNA-binding proteins. *Wiley Interdiscip. Rev. RNA* **5**, 111–130 (2014).

113. Nitta, K. R. *et al.* Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **4**, (2015).

114. Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–3182 (2009).

115. Liefooghe, A., Touzet, H. & Varré, J.-S. Large Scale Matching for Position Weight Matrices. in *Combinatorial Pattern Matching* 401–412 (Springer Berlin Heidelberg, 2006).

116. Ji, H. *et al.* An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* **26**, 1293–1300 (2008).

117. Paz, I., Kosti, I., Ares, M., Jr, Cline, M. & Mandel-Gutfreund, Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.* **42**, W361–7 (2014).

118. Beckstette, M., Homann, R., Giegerich, R. & Kurtz, S. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* **7**, 389 (2006).

119. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).

120. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

121. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation.

Am. J. Hum. Genet. **101**, 5–22 (2017).

122. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).

123. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).

124. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).

125. Sun, W. & Hu, Y. eQTL Mapping Using RNA-seq Data. *Stat. Biosci.* **5**, 198–219 (2013).

126. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25–33 (2000).

127. Urnov, F. D., Rebar, E. J., Holmes, M. C., Steve Zhang, H. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nature Reviews Genetics* **11**, 636–646 (2010).

128. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509–1512 (2009).

129. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29**, 143–148 (2011).

130. Maher, B. ENCODE: The human encyclopaedia. *Nature* **489**, 46–48 (2012).

131. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806–811 (1998).

132. Mohr, S., Bakal, C. & Perrimon, N. Genomic screening with RNAi: results and challenges. *Annu. Rev. Biochem.* **79**, 37–64 (2010).

133. Fedorov, Y. *et al.* Off-target effects by siRNA can induce toxic phenotype. *RNA* **12**, 1188–1196 (2006).

134. Jackson, A. L. *et al.* Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol.* **21**, 635–637 (2003).

135. Boutros, M. & Ahringer, J. The art and design of genetic screens: RNA interference. *Nat. Rev. Genet.* **9**, 554–566 (2008).

136. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32**, 347–355 (2014).

137. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).

138. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–661 (2014).

139. Doench, J. G. Am I ready for CRISPR? A user's guide to genetic screens. *Nat. Rev. Genet.* **19**, 67–80 (2018).

140. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1911 (2018).

141. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

142. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).

143. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).

144. Zhang, X. D. Illustration of SSMD, z score, SSMD*, z* score, and t statistic for hit selection in RNAi high-throughput screens. *J. Biomol. Screen.* **16**, 775–785 (2011).

145. Amundadottir, L. T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658 (2006).

146. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).

147. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).

148. Al Olama, A. A. *et al.* Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat. Genet.* **41**, 1058–1060 (2009).

149. Yeager, M. *et al.* Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **41**, 1055–1057 (2009).

150. Gao, L. *et al.* Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nat. Commun.* **9**, 702 (2018).

151. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–1543 (2015).

152. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).

153. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of

human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

154. Tsuda, Y., Nishida-Umehara, C., Ishijima, J., Yamada, K. & Matsuda, Y. Comparison of the Z and W sex chromosomal architectures in elegant crested tinamou (Eudromia elegans) and ostrich (Struthio camelus) and the process of sex chromosome differentiation in palaeognathous birds. *Chromosoma* **116**, 159–173 (2007).

155. Zhang, G. *et al.* Comparative Genomics Across Modern Bird Species Reveal Insights into Panavian Genome evolution and Trait Biodiversity. *Science* 1311–1320.

156. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).

157. Ottolenghi, C., Fellous, M., Barbieri, M. & McElreavey, K. Novel paralogy relations among human chromosomes support a link between the phylogeny of doublesex-related genes and the evolution of sex determination. *Genomics* **79**, 333–343 (2002).

158. Brunner, B. *et al.* Genomic organization and expression of the doublesex-related gene cluster in vertebrates and detection of putative regulatory regions for DMRT1. *Genomics* **77**, 8–17 (2001).

159. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* (2009).

160. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

161. Fan, W.-L. *et al.* Genome-wide patterns of genetic variation in two domestic chickens. *Genome Biol. Evol.* **5**, 1376–1392 (2013).

162. Vicoso, B., Kaiser, V. B. & Bachtrog, D. Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 6453–6458 (2013).

163. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* (2002).

164. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).

165. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

166. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).

167. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

168. Adolfsson, S. & Ellegren, H. Lack of dosage compensation accompanies the arrested stage of sex chromosome evolution in ostriches. *Mol. Biol. Evol.* **30**, 806–810 (2013).

169. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

170. Alföldi, J. *et al.* The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**, 587–591 (2011).

171. Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).

172. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).

173. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).

174. Loytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences* **102**, 10557–10562 (2005).

175. Zhou, Q. *et al.* Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346**, 1246338 (2014).

176. Lamesch, P. *et al.* hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics* **89**, 307–315 (2007).

177. Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol. Cell* **70**, 854–867.e9 (2018).

178. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* **39**, D301–8 (2011).

179. Farazi, T. A. *et al.* Identification of the RNA recognition element of the RBPMS family of RNA-binding proteins and their transcriptome-wide mRNA targets. *RNA* **20**, 1090–1102 (2014).

180. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).

181. Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.* **6**, 1139–1152 (2014).

182. Haque, N., Ouda, R., Chen, C., Ozato, K. & Hogg, J. R. ZFR coordinates crosstalk between RNA decay and transcription in innate immunity. *Nat. Commun.* **9**, 1145 (2018).

183. Huang, H. *et al.* Recognition of RNA N6-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nat. Cell Biol.* **20**, 285–295 (2018).

184. Damianov, A., Kann, M., Lane, W. S. & Bindereif, A. Human RBM28 protein is a specific nucleolar component of the spliceosomal snRNPs. *Biol. Chem.* **387**, 1455–1460 (2006).

185. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–62 (2009).

186. Zhang, X. D. *et al.* The use of strictly standardized mean difference for hit selection in primary RNA interference high-throughput screening experiments. *J. Biomol. Screen.* **12**, 497–509 (2007).

187. Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).

188. Dave, K. *et al.* Mice deficient of Myc super-enhancer region reveal differential control mechanism between normal and pathological growth. *Elife* **6**, (2017).

189. Lemaitre, C. *et al.* Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. *Genome Biol. Evol.* **1**, 56–66 (2009).

190. Rice, W. R. The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination between Primitive Sex Chromosomes. *Evolution* **41**, 911–914 (1987).

191. Charlesworth, D. & Charlesworth, B. Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. *Genet. Res.* **35**, 205–214 (1980).

192. Varani, G. & McClain, W. H. The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep.* **1**, 18–23 (2000).

193. Loughlin, F. E. *et al.* The zinc fingers of the SR-like protein ZRANB2 are single-stranded RNA-binding domains that recognize 5' splice site-like sequences. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 5581–5586 (2009).

194. Lee, H. *et al.* Zinc-finger antiviral protein mediates retinoic acid inducible gene I-like receptor-independent antiviral response to murine leukemia virus. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 12379–12384 (2013).